

# TRANSPORTATION RESEARCH RECORD

---

Journal of the Transportation Research Board, No. 2289

Railways  
2012

**TRANSPORTATION RESEARCH RECORDS**, which are published throughout the year, consist of collections of papers on specific transportation modes and subject areas. Each Record is classified according to the subscriber category or categories covered by the papers published in that volume. The views expressed in papers published in the Transportation Research Record series are those of the authors and do not necessarily reflect the views of the peer review committee(s), the Transportation Research Board, the National Research Council, or the sponsors of TRB activities. The Transportation Research Board does not endorse products or manufacturers; trade and manufacturers' names may appear in a Record paper only if they are considered essential.

**PEER REVIEW OF PAPERS:** All papers published in the Transportation Research Record series have been reviewed and accepted for publication through the Transportation Research Board's peer review process established according to procedures approved by the Governing Board of the National Research Council. Papers are refereed by the TRB standing committees identified on page ii of each Record. Reviewers are selected from among committee members and other outside experts. The Transportation Research Board requires a minimum of three reviews; a decision is based on reviewer comments and resultant author revision. For details about the peer review process, see the information on the inside back cover.

**THE TRANSPORTATION RESEARCH RECORD PUBLICATION BOARD**, comprising a cross section of transportation disciplines and with equal representation from the academic and practitioner communities, assures that the quality of the *Transportation Research Record: Journal of the Transportation Research Board* and the TRB paper peer review process is consistent with that of a peer-reviewed scientific journal. Members from the academic community make decisions on granting tenure; all academic and practitioner members have authored papers published in peer-reviewed journals and all have participated in the TRB peer review process.

**TRANSPORTATION RESEARCH RECORD PAPERS ONLINE:** The TRR Journal Online website provides electronic access to the full text of papers that have been published in the Transportation Research Record series since 1996. The site is updated as new Record papers become available. To search abstracts and to find subscription and pricing information, go to [www.TRB.org/TRROnline](http://www.TRB.org/TRROnline).

**TRANSPORTATION RESEARCH BOARD PUBLICATIONS** may be ordered directly from the TRB Business Office, through the Internet at [www.TRB.org](http://www.TRB.org), or by annual subscription through organizational or individual affiliation with TRB. Affiliates and library subscribers are eligible for substantial discounts. For further information, contact the Transportation Research Board Business Office, 500 Fifth Street, NW, Washington, DC 20001 (telephone 202-334-3213; fax 202-334-2519; or e-mail [TRBSales@nas.edu](mailto:TRBSales@nas.edu)).

## TRANSPORTATION RESEARCH BOARD 2012 EXECUTIVE COMMITTEE\*

**Chair:** Sandra Rosenbloom, Professor of Planning, University of Arizona, Tucson  
**Vice Chair:** Deborah H. Butler, Executive Vice President, Planning, and CIO, Norfolk Southern Corporation, Norfolk, Virginia  
**Executive Director:** Robert E. Skinner, Jr., Transportation Research Board

**Victoria A. Arroyo**, Executive Director, Georgetown Climate Center, and Visiting Professor, Georgetown University Law Center, Washington, D.C.  
**J. Barry Barker**, Executive Director, Transit Authority of River City, Louisville, Kentucky  
**William A. V. Clark**, Professor of Geography (emeritus) and Professor of Statistics (emeritus), Department of Geography, University of California, Los Angeles  
**Eugene A. Conti, Jr.**, Secretary of Transportation, North Carolina Department of Transportation, Raleigh  
**James M. Crites**, Executive Vice President of Operations, Dallas-Fort Worth International Airport, Texas  
**Paula J. C. Hammond**, Secretary, Washington State Department of Transportation, Olympia  
**Michael W. Hancock**, Secretary, Kentucky Transportation Cabinet, Frankfort  
**Chris T. Hendrickson**, Duquesne Light Professor of Engineering, Carnegie Mellon University, Pittsburgh, Pennsylvania  
**Adib K. Kanafani**, Professor of the Graduate School, University of California, Berkeley (Past Chair, 2009)  
**Gary P. LaGrange**, President and CEO, Port of New Orleans, Louisiana  
**Michael P. Lewis**, Director, Rhode Island Department of Transportation, Providence  
**Susan Martinovich**, Director, Nevada Department of Transportation, Carson City  
**Joan McDonald**, Commissioner, New York State Department of Transportation, Albany  
**Michael R. Morris**, Director of Transportation, North Central Texas Council of Governments, Arlington (Past Chair, 2010)  
**Tracy L. Rosser**, Vice President, Regional General Manager, Wal-Mart Stores, Inc., Mandeville, Louisiana  
**Henry G. (Gerry) Schwartz, Jr.**, Chairman (retired), Jacobs/Sverdrup Civil, Inc., St. Louis, Missouri  
**Beverly A. Scott**, General Manager and CEO, Metropolitan Atlanta Rapid Transit Authority, Atlanta, Georgia  
**David Seltzer**, Principal, Mercator Advisors LLC, Philadelphia, Pennsylvania  
**Kumares C. Sinha**, Olson Distinguished Professor of Civil Engineering, Purdue University, West Lafayette, Indiana  
**Thomas K. Sorel**, Commissioner, Minnesota Department of Transportation, St. Paul  
**Daniel Sperling**, Professor of Civil Engineering and Environmental Science and Policy; Director, Institute of Transportation Studies; and Acting Director, Energy Efficiency Center, University of California, Davis  
**Kirk T. Steudle**, Director, Michigan Department of Transportation, Lansing  
**Douglas W. Stotlar**, President and Chief Executive Officer, Con-Way, Inc., Ann Arbor, Michigan  
**C. Michael Walton**, Ernest H. Cockrell Centennial Chair in Engineering, University of Texas, Austin (Past Chair, 1991)  
**Rebecca M. Brewster**, President and COO, American Transportation Research Institute, Smyrna, Georgia (ex officio)  
**Anne S. Ferro**, Administrator, Federal Motor Carrier Safety Administration, U.S. Department of Transportation (ex officio)  
**LeRoy Gishi**, Chief, Division of Transportation, Bureau of Indian Affairs, U.S. Department of the Interior, Washington, D.C. (ex officio)  
**John T. Gray II**, Senior Vice President, Policy and Economics, Association of American Railroads, Washington, D.C. (ex officio)  
**John C. Horsley**, Executive Director, American Association of State Highway and Transportation Officials, Washington, D.C. (ex officio)  
**Michael P. Huerta**, Acting Administrator, Federal Aviation Administration, U.S. Department of Transportation (ex officio)  
**David T. Matsuda**, Administrator, Maritime Administration, U.S. Department of Transportation (ex officio)  
**Michael P. Melaniphy**, President and CEO, American Public Transportation Association, Washington, D.C. (ex officio)  
**Victor M. Mendez**, Administrator, Federal Highway Administration, U.S. Department of Transportation (ex officio)  
**Tara O'Toole**, Under Secretary for Science and Technology, U.S. Department of Homeland Security (ex officio)  
**Robert J. Papp** (Adm., U.S. Coast Guard), Commandant, U.S. Coast Guard, U.S. Department of Homeland Security (ex officio)  
**Cynthia L. Quarterman**, Administrator, Pipeline and Hazardous Materials Safety Administration, U.S. Department of Transportation (ex officio)  
**Peter M. Rogoff**, Administrator, Federal Transit Administration, U.S. Department of Transportation (ex officio)  
**David L. Strickland**, Administrator, National Highway Traffic Safety Administration, U.S. Department of Transportation (ex officio)  
**Joseph C. Szabo**, Administrator, Federal Railroad Administration, U.S. Department of Transportation (ex officio)  
**Polly Trottenberg**, Assistant Secretary for Transportation Policy, U.S. Department of Transportation (ex officio)  
**Robert L. Van Antwerp** (Lt. General, U.S. Army), Chief of Engineers and Commanding General, U.S. Army Corps of Engineers, Washington, D.C. (ex officio)  
**Barry R. Wallerstein**, Executive Officer, South Coast Air Quality Management District, Diamond Bar, California (ex officio)  
**Gregory D. Winfree**, Acting Administrator, Research and Innovative Technology Administration, U.S. Department of Transportation (ex officio)

\* Membership as of October 2012.



# **TRANSPORTATION RESEARCH RECORD**

---

Journal of the Transportation Research Board, No. 2289

**Railways  
2012**

A Peer-Reviewed Publication

---

**TRANSPORTATION RESEARCH BOARD**  
*OF THE NATIONAL ACADEMIES*

Washington, D.C.  
2012

[www.TRB.org](http://www.TRB.org)

## Transportation Research Record 2289

ISSN 0361-1981 (print)

ISSN 2169-4052 (online)

ISBN 978-0-309-22329-4

### Subscriber Categories

Rail; freight transportation; passenger transportation

Printed in the United States of America

## TRANSPORTATION RESEARCH RECORD PUBLICATION BOARD

C. Michael Walton, PhD, PE, Ernest H. Cockrell Centennial Chair in Engineering and Professor of Civil Engineering, University of Texas, Austin (Cochair)

Mary Lynn Tischer, PhD, Director, Office of Transportation Policy Studies, Federal Highway Administration, Washington, D.C. (Cochair)

Daniel Brand, SB, SM, PE, Consultant, Lyme, New Hampshire

Mary R. Brooks, BOT, MBA, PhD, William A. Black Chair of Commerce, Dalhousie University, Halifax, Nova Scotia, Canada

Charles E. Howard, Jr., MCRP, Director, Transportation Planning, Puget Sound Regional Council, Seattle, Washington

Thomas J. Kazmierowski, PEng, Manager, Materials Engineering and Research Office, Ontario Ministry of Transportation, Toronto, Canada

Michael D. Meyer, PhD, PE, Frederick R. Dickerson Professor, School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta

Sandra Rosenbloom, MA, PhD, Professor of Planning, University of Arizona, Tucson

Kumares C. Sinha, PhD, PE, Olson Distinguished Professor, School of Civil Engineering, Purdue University, West Lafayette, Indiana

L. David Suits, MSCE, Executive Director, North American Geosynthetics Society, Albany, New York

## Peer Review of Transportation Research Record 2289

### RAIL GROUP

Anthony D. Perl, Simon Fraser University (Chair)

### Intercity Passenger Rail Committee

David P. Simpson, David P. Simpson Consultants, LLC (Chair), Camille Tsao, HNTB Corporation (Secretary), Rohit T. Aggarwala, Daniel Brand, Ross B. Capon, Rod J. Diridon, Penny E. Eickemeyer, Moshe Givoni, Chelsea L. Gleis, Dharm Guruswamy, George Haikalis, Olivier Klein, Felix Laube, Alexander Lu, Jason J. Maga, Deborah Wood Matherly, Ronald A. Mauri, Matthew James Melzer, Curtis A. Morgan, Andrew B. Nash, Anthony D. Perl, Howard R. Permut, Eric C. Peterson, Michael Angel Rodriguez, Daniel L. Roth, Allan Rutter, Michael Huntly Schabas, Reed H. Tanger, John C. Tone, Eric S. Tyrer, II, Randall E. Wade, Mark C. Walbrun

### Passenger Rail Equipment and Systems Integration Committee

Patrick B. Simmons, North Carolina Department of Transportation (Chair), Eloy E. Martinez, LTK Engineering Services, Inc. (Secretary), John A. Boffa, Joshua Coran, Robert M. Dorer, Chad R. Edison, John A. Harrison, Steven J. Hewitt, Suzanne M. Horton, Stanton C. Hunter, Katharine M. Hunter-Zaworski, J. Lee Hutchins, Jr., Filippo Martinelli, Rodney P. Massman, David O. Nelson, Charles A. Poltenson, Sr., Katie Stanchak, Michael J. Trosino, David A. Valenstein, Davidson A. Ward

### Railroad Operating Technologies Committee

Yung-Cheng Lai, National Taiwan University (Chair), Adrian D. Hellman, Research and Innovative Technology Administration (Secretary), Sam S. Alibrahim, Christopher P. L. Barkan, Virginia M. Beck, Stephen Juan Bruno, Olga K. Cataldi, David B. Clarke, Richard U. Cogswell, Llewellyn C. Davis, Timothy J. DePaepe, Mark H. Dingler, Mark W. Hemphill, E. Keith Holt, Michael Iden, Edwin R. Kraft, Coleman Lawrence, Robert H. Leilich, Marco M. Luethi, Larry R. Milhon, Jena C. Montgomery, Gordon B. Mott, Joern Pacht, Paul H. Reistrup, Mark K. Ricci, Ismail Sahin, Victor Simuoli, Michael E. Smith, Mark P. Stehly, James A. Stem, Jr.

## Freight Rail Transportation Committee

George Avery Grimes, Metra (Chair), Joe L. Arbona, Christopher P. L. Barkan, David B. Clarke, Diane Davidson, John T. Gray, II, J. Scott Greene, Aaron P. Hegeman, Pasi T. Lautala, Edward A. Lewis, Donald B. Ludlow, David H. Mangold, Tamara L. Nicholson, Elizabeth E. Ogard, Henry Posner, III, Steven A. Potter, George Raymond, Randolph R. Resor, James H. M. Savage, George W. Schafer, III, David P. Simpson, Chris S. Smith, Peter F. Swan, Forrest R. Van Schwartz, Jerry Ellison Vest, Jr.

## Railroad Track Structure System Design Committee

David D. Davis, Transportation Technology Center, Inc. (Chair), Carlton L. Ho, University of Massachusetts, Amherst (Secretary), Ernest J. Barenberg, Timothy R. Bennett, David N. Bilow, Randy L. Bowman, Michael O. Brown, Miodrag Budisa, William G. Byers, Dwight W. Clark, Laurence E. Daniels, Riley Edwards, Konstantinos Giannakos, Hannes Grabe, Henry M. Lees, Jr., Mohammad S. Longi, Andrés López-Pita, Dale W. Ophardt, Donald Plotkin, Juanjuan Ren, David E. Staplin, Yu-Jiang Zhang

## Rail Transit Infrastructure Committee

Bruce R. Smith, Gannett Fleming Transit and Rail Systems (Chair), Steven Abramopoulos, David A. Boate, Anthony P. Bohara, Michael O. Brown, Stelian Canjea, B. N. Reddy Chidananda, Laurence E. Daniels, Hugh J. Fuller, Peter S. Gentle, Brent Graham, Lawrence G. Lovejoy, William H. Moorhead, Paul G. Pattison, Randal S. Phelan, Hugh Saurenman, Pranaya Shrestha, Philip M. Strong, Peter Torres, Eduardo Ugarte, Norman Vutz, John F. Zuspan

## Railway Maintenance Committee

Dwight W. Clark, Union Pacific Railroad Company (Chair), Hai Huang, Pennsylvania State University (Secretary), Miodrag Budisa, Gary A. Carr, Thomas F. DeJoseph, Marcus S. Dersch, Riley Edwards, Gurmel S. Ghataora, Konstantinos Giannakos, Hannes Grabe, Carlton L. Ho, Francesco Lanza Di Scalea, Michael McMasters, Dennis W. Morgart, Thomas Henry O'Brien, Bruce R. Pohlott, Hamed Pouryousef, Christian Mark Roberts, Henry A. Rubert, Mario A. Ruel, Eric Sherrock, Jackie van der Westhuizen, Timothy R. Wells

## Railroad Operational Safety Committee

Stephen M. Popkin, Research and Innovative Technology Administration (Chair), Ann M. Mills, Rail Safety and Standards Board (Vice Chair), Jacquelyn M. Keenan, Union Pacific Railroad Company (Secretary), Gina M. Melnik, Research and Innovative Technology Administration (Secretary), Michael K. Coplen, Grady C. Cothen, Jr., Timothy J. DePaepe, Jason Travis Doering, Lawrence B. Fleischer, Judith B. Gertler, Peter D. Hall, Vijay K. Kohli, Jennifer E. Lincoln, David H. Mangold, Jeffrey Franklin Moller, Charles M. Oman, Thomas A. Pontolillo, Thomas G. Raslear, Mark K. Ricci, Derrell Ross, Patrick Sherry, James A. Stem, Jr., Thomas E. Streicher

Peer review is indicated by a footnote at the end of each paper. The organizational units, officers, and members are as of December 31, 2011.

## Transportation Research Board Staff

Ann R. Purdue, Senior Program Officer, Rail and Freight  
Matthew A. Miller, Senior Program Associate  
Richard F. Pain, Senior Program Officer and Transportation Safety Coordinator  
Freda R. Morgan, Senior Program Associate  
Mary Kissi, Senior Program Associate

### Publications Office

Diane LeBlanc Solometo, Editor; Benjamin R. Justesen, Production Editor; Mary McLaughlin, Proofreader; Peter Dull, Manuscript Preparer  
Ann E. Petty, Managing Editor; Juanita Green, Production Manager; Phyllis Barber, Publishing Services Manager; Jennifer J. Weeks, Manuscript Preparation Manager; Jennifer Corroero, Senior Editorial Assistant; Paul deBruijn, Production Assistant

# THE NATIONAL ACADEMIES

## *Advisers to the Nation on Science, Engineering, and Medicine*

The **National Academy of Sciences** is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. On the authority of the charter granted to it by the Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Ralph J. Cicerone is president of the National Academy of Sciences.

The **National Academy of Engineering** was established in 1964, under the charter of the National Academy of Sciences, as a parallel organization of outstanding engineers. It is autonomous in its administration and in the selection of its members, sharing with the National Academy of Sciences the responsibility for advising the federal government. The National Academy of Engineering also sponsors engineering programs aimed at meeting national needs, encourages education and research, and recognizes the superior achievements of engineers. Dr. Charles M. Vest is president of the National Academy of Engineering.

The **Institute of Medicine** was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, on its own initiative, to identify issues of medical care, research, and education. Dr. Harvey V. Fineberg is president of the Institute of Medicine.

The **National Research Council** was organized by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and advising the federal government. Functioning in accordance with general policies determined by the Academy, the Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities. The Council is administered jointly by both Academies and the Institute of Medicine. Dr. Ralph J. Cicerone and Dr. Charles M. Vest are chair and vice chair, respectively, of the National Research Council.

The **Transportation Research Board** is one of six major divisions of the National Research Council. The mission of the Transportation Research Board is to provide leadership in transportation innovation and progress through research and information exchange, conducted within a setting that is objective, interdisciplinary, and multimodal. The Board's varied activities annually engage about 7,000 engineers, scientists, and other transportation researchers and practitioners from the public and private sectors and academia, all of whom contribute their expertise in the public interest. The program is supported by state transportation departments, federal agencies including the component administrations of the U.S. Department of Transportation, and other organizations and individuals interested in the development of transportation.

[www.TRB.org](http://www.TRB.org)

[www.national-academies.org](http://www.national-academies.org)

## TRB SPONSORS\*

Transportation Departments of the 50 States and the District of Columbia

### *Federal Government*

U.S. Department of Transportation

Federal Aviation Administration

Federal Highway Administration

Federal Motor Carrier Safety Administration

Federal Railroad Administration

Federal Transit Administration

National Highway Traffic Safety Administration

Research and Innovative Technology Administration

Bureau of Indian Affairs

Office of Naval Research, U.S. Navy

Science and Technology Directorate,

U.S. Department of Homeland Security

U.S. Army Corps of Engineers

U.S. Coast Guard

### *Nongovernmental Organizations*

American Association of State Highway  
and Transportation Officials

American Public Transportation Association

American Transportation Research Institute

Association of American Railroads

South Coast Air Quality Management District, California

---

\*As of October 2012.

# TRANSPORTATION RESEARCH RECORD

Journal of the Transportation Research Board, No. 2289

## Contents

<b>Foreword</b>	<b>vii</b>
<b>Assessment of Advanced Dispatching Measures for Recovering Disrupted Railway Traffic Situations</b> Francesco Corman and Andrea D'Ariano	<b>1</b>
<b>High-Speed Rail Versus Air Transportation: Case Study of Madrid-Barcelona, Spain</b> Francesca Pagliara, José Manuel Vassallo, and Concepción Román	<b>10</b>
<b>High-Speed Route Improvement Optimizer</b> Yung-Cheng (Rex) Lai and Po-Wen Huang	<b>18</b>
<b>Decision Support System to Optimize Railway Stopping Patterns: Application to Taiwan High-Speed Rail</b> Jyh-Cherng Jong, Chian-Shan (James) Suen, and S. K. (Jason) Chang	<b>24</b>
<b>Risk Assessment of Positive Train Control by Using Simulation of Rare Events</b> Timothy Meyers, Amine Stambouli, Karen McClure, and Daniel Brod	<b>34</b>
<b>Dual-Mode and New Diesel Locomotive Developments</b> Janis Vitins	<b>42</b>
<b>Development of the Next Generation of Intercity Corridor Bi-Level Equipment with Crash Energy Management</b> Eloy Martinez, Frances Nelson, Anand Prabhakaran, and Antony Jones	<b>47</b>
<b>Portable Emission Measurement System for Emissions of Passenger Rail Locomotives</b> H. Christopher Frey, Hyung-Wook Choi, and Kangwook Kim	<b>56</b>



<b>Slab Track Mass-Spring System</b> Mirjana Tomicic-Torlakovic, Miodrag Budisa, and Vidan Radjen	<b>64</b>
<b>Comparison of Magnitude of Actions on Track in High-Speed and Heavy-Haul Railroads: Influence of Resilient Fastenings</b> Konstantinos Giannakos	<b>70</b>
<b>Movement of Water Through Ballast and Subballast for Dual-Line Railway Track</b> Gurmeh S. Ghataora and Ken Rushton	<b>78</b>
<b>Source of Ballast Fouling and Influence Considerations for Condition Assessment Criteria</b> Ted R. Sussmann, Mario Ruel, and Steven M. Chrismer	<b>87</b>
<b>Ground-Penetrating Radar Data to Develop Wavelet Technique for Quantifying Railroad Ballast-Fouling Conditions</b> Pengcheng Shangguan, Imad L. Al-Qadi, and Zhen Leng	<b>95</b>
<b>Stochastic Rail Wear Model for Railroad Tracks</b> Seosamh B. Costello, Anuradha S. Premathilaka, and Roger C. M. Dunn	<b>103</b>
<b>Asset Condition Assessment at Regional Transportation Authority in Chicago, Illinois</b> Grace Gallucci, John Goodworth, and John G. Allen	<b>111</b>
<b>Development of Base Train Equivalents to Standardize Trains for Capacity Analysis</b> Yung-Cheng (Rex) Lai, Yun-Hsuan Liu, and Tzu-Ya Lin	<b>119</b>
<b>Methodological Framework for Analyzing Ability of Freight Rail Customers to Forecast Short-Term Volumes Accurately</b> Stephan Moll, Ulrich Weidmann, and Andrew Nash	<b>126</b>
<b>Value for Railway Capacity: Assessing Efficiency of Operators in Great Britain</b> Melody Khadem Sameni and John M. Preston	<b>134</b>
<b>Development and Assessment of Taxonomy for Performance-Shaping Factors for Railway Operations</b> Miltos Kyriakidis, Arnab Majumdar, Gudela Grote, and Washington Y. Ochieng	<b>145</b>
<b>Analysis of Causes of Major Train Derailment and Their Effect on Accident Rates</b> Xiang Liu, M. Rapik Saat, and Christopher P. L. Barkan	<b>154</b>

# Foreword

The 2012 series of the *Transportation Research Record: Journal of the Transportation Research Board* consists of approximately 940 papers selected from more than 4,000 submissions after rigorous peer review. The peer review for each paper published in this volume was coordinated by the committee acknowledged at the end of the text; members of the reviewing committees for the papers in this volume are listed on page ii.

Additional information about the *Transportation Research Record: Journal of the Transportation Research Board* series and the peer review process appears on the inside back cover. TRB appreciates the interest shown by authors in offering their papers, and the Board looks forward to future submissions.

**Note:** Many of the photographs, figures, and tables in this volume have been converted from color to grayscale for printing. The electronic files of the papers, posted on the web at **[www.TRB.org/TRROnline](http://www.TRB.org/TRROnline)**, retain the color versions of photographs, figures, and tables as originally submitted for publication.

## Measurement Conversion Factors

To convert from the unit in the first column to the unit in the second column, multiply by the factor in the third column.

<i>Customary Unit</i>	<i>SI Unit</i>	<i>Factor</i>
<b>Length</b>		
inches	millimeters	25.4
inches	centimeters	2.54
feet	meters	0.305
yards	meters	0.914
miles	kilometers	1.61
<b>Area</b>		
square inches	square millimeters	645.1
square feet	square meters	0.093
square yards	square meters	0.836
acres	hectares	0.405
square miles	square kilometers	2.59
<b>Volume</b>		
gallons	liters	3.785
cubic feet	cubic meters	0.028
cubic yards	cubic meters	0.765
<b>Mass</b>		
ounces	grams	28.35
pounds	kilograms	0.454
short tons	megagrams	0.907
<b>Illumination</b>		
footcandles	lux	10.76
footlamberts	candelas per square meter	3.426
<b>Force and Pressure or Stress</b>		
poundforce	newtons	4.45
poundforce per square inch	kilopascals	6.89
<b>Temperature</b>		

To convert Fahrenheit temperature (°F) to Celsius temperature (°C), use the following formula:  
 $^{\circ}\text{C} = (^{\circ}\text{F} - 32)/1.8$

<i>SI Unit</i>	<i>Customary Unit</i>	<i>Factor</i>
<b>Length</b>		
millimeters	inches	0.039
centimeters	inches	0.394
meters	feet	3.281
meters	yards	1.094
kilometers	miles	0.621
<b>Area</b>		
square millimeters	square inches	0.00155
square meters	square feet	10.764
square meters	square yards	1.196
hectares	acres	2.471
square kilometers	square miles	0.386
<b>Volume</b>		
liters	gallons	0.264
cubic meters	cubic feet	35.314
cubic meters	cubic yards	1.308
<b>Mass</b>		
grams	ounces	0.035
kilograms	pounds	2.205
megagrams	short tons	1.102
<b>Illumination</b>		
lux	footcandles	0.093
candelas per square meter	footlamberts	0.292
<b>Force and Pressure or Stress</b>		
newtons	poundforce	0.225
kilopascals	poundforce per square inch	0.145
<b>Temperature</b>		

To convert Celsius temperature (°C) to Fahrenheit temperature (°F), use the following formula:  
 $^{\circ}\text{F} = (^{\circ}\text{C} \times 1.8) + 32$

## Abbreviations Used Without Definitions

AASHO	American Association of State Highway Officials
AASHTO	American Association of State Highway and Transportation Officials
ACRP	Airport Cooperative Research Program
APTA	American Public Transportation Association
ASCE	American Society of Civil Engineers
ASTM	American Society for Testing and Materials (known by abbreviation only)
FAA	Federal Aviation Administration
FHWA	Federal Highway Administration
FMCSA	Federal Motor Carrier Safety Administration
FRA	Federal Railroad Administration
FTA	Federal Transit Administration
IEEE	Institute of Electrical and Electronics Engineers
ISO	International Organization for Standardization
ITE	Institute of Transportation Engineers
NASA	National Aeronautics and Space Administration
NCHRP	National Cooperative Highway Research Program
NHTSA	National Highway Traffic Safety Administration
RITA	Research and Innovative Technology Administration
SAE	Society of Automotive Engineers
SHRP	Strategic Highway Research Program
TCRP	Transit Cooperative Research Program
TRB	Transportation Research Board

# Assessment of Advanced Dispatching Measures for Recovering Disrupted Railway Traffic Situations

Francesco Corman and Andrea D'Ariano

Railway timetables are developed to make operations robust and resilient to small delays. However, disturbances perturb the daily plan, and dispatchers need to adjust the plan to keep operations feasible and to limit delay propagation. For large infrastructure disruptions, the available railway capacity is reduced, and the timetable could become infeasible. The paper studies how to support dispatchers in the management of traffic flow during disruptions. A set of disruption resolution scenarios to manage seriously disturbed traffic conditions in large networks is investigated. For instance, in the case of track blockage, train services can be canceled, rerouted in the disrupted dispatching area, or rerouted in other areas while still with the same origin and destination. Feasible and efficient operations schedules are found quickly by an advanced decision support system for dispatching known as ROMA (railway traffic optimization by means of alternative graphs), which is based on microscopic detail and can handle large areas by decomposition. Detailed performance indicators can be computed to let dispatchers choose a specific solution, for example, by minimizing train delays and reducing passengers' discomfort. In the computational experiments, an analysis is done of a blockage on a double track line, combined with multiple entrance delays on a large railway network with heavy traffic. Several disruption resolution scenarios involving cancellation of services, rerouting, and shuttle trains are considered, and each feasible plan is evaluated in relation to travel times, frequency of services, and delay propagation.

The continuous growth in the frequency of passenger and freight railway traffic is increasing the pressure on railway companies. Improvement of the reliability of rail transit operations is based on the design of robust timetables that should be able to deal with minor perturbations (i.e., a few minutes of delays) occurring in real time by using smart planning rules and time reserves. However, no reasonable railway plan is robust or reliable enough in the case of technical failures and other disturbances (such as large train delays, reduced operating speeds, bad weather, and temporary unavailability of some routes). Those disrupted traffic situations may seriously influence the running times, dwelling, and departing events of trains, thus causing nonscheduled waiting times and longer running times (1–3).

---

F. Corman, Department of Mechanical Engineering, Section Traffic and Infrastructure, Katholieke Universiteit Leuven, Celestijnenlaan 300A–P.O. Box 2422, Heverlee 3001, Belgium. A. D'Ariano, Dipartimento di Informatica e Automazione, Università Degli Studi Roma Tre, Via Della Vasca Navale 79, Rome 00146, Italy. Corresponding author: F. Corman, francesco.corman@cib.kuleuven.be.

*Transportation Research Record: Journal of the Transportation Research Board*, No. 2289, Transportation Research Board of the National Academies, Washington, D.C., 2012, pp. 1–9.  
DOI: 10.3141/2289-01

As a result of the interaction between trains, such disturbances can propagate as knock-on delays to other trains in the network. During the disruption, dispatchers need to alter the plan with regard to train travel times, orders, and routes required to recover the feasibility of disrupted operations, under the shortage of spare railway capacity and several delayed trains.

Operational traffic management in large and busy networks is typically based on train plans that are defined offline, even in the presence of blocked tracks. In the Netherlands, so-called emergency timetables are used as a response to disrupted operations (formally defined as some track blocked for more than 30 min and causing trains to be delayed by more than 5 min). Emergency timetables are defined to cope with all possible infrastructure malfunctions and further grouped in disruptions of similar characteristics to limit the total number of cases considered (1,200 emergency timetables are currently considered).

Figure 1 shows an example of an emergency timetable. The disruption under discussion is reported graphically in regard to the infrastructure unavailability, namely a block on one of the two tracks from Dordrecht and Lage Zwaluwe (left) to Zevenbergen and Roosendaal (right). Figure 1 describes the situation—one of the two tracks between Lage Zwaluwe and Zevenbergen is blocked—and the modifications to passenger services (Reizigerstreinen) in relation to line number, details, and control region. The modifications to passenger services are in the area named Randstad South and are as follows: Services 600 and 9300 are kept running (blijft rijden), while Line Series 2100 and 5100 are canceled (opheffen) between the stations of Dordrecht and Roosendaal, and Lage Zwaluwe and Oudenbosch, respectively.

Optimization methodologies to support dispatchers in developing operations strategies for responding to major traffic disruptions on busy railways are studied. Managing disrupted traffic is one of the most challenging dispatching tasks on railroads with high-density passenger traffic to be operated on reduced infrastructure. Experienced traffic controllers have developed strategies allowing them simply to foresee possible disruptions well in advance and to take compensatory control actions on the basis of local information. Dispatchers reschedule the route setting plan only when trains experience a considerable delay. However, during exceptional situations their experience and planned emergency timetables cannot support them in understanding the complete consequences of the intended rescheduling actions.

Decision support tools are thus needed to help dispatchers manage disruptions. The solution presented in this paper is based on the idea that the computer evaluates several predefined scenarios to make a decision about which of these scenarios would be the most favorable.

## Maatregel 06H-A

Geldig vanaf: 10 dec 2006

Geldig tot:

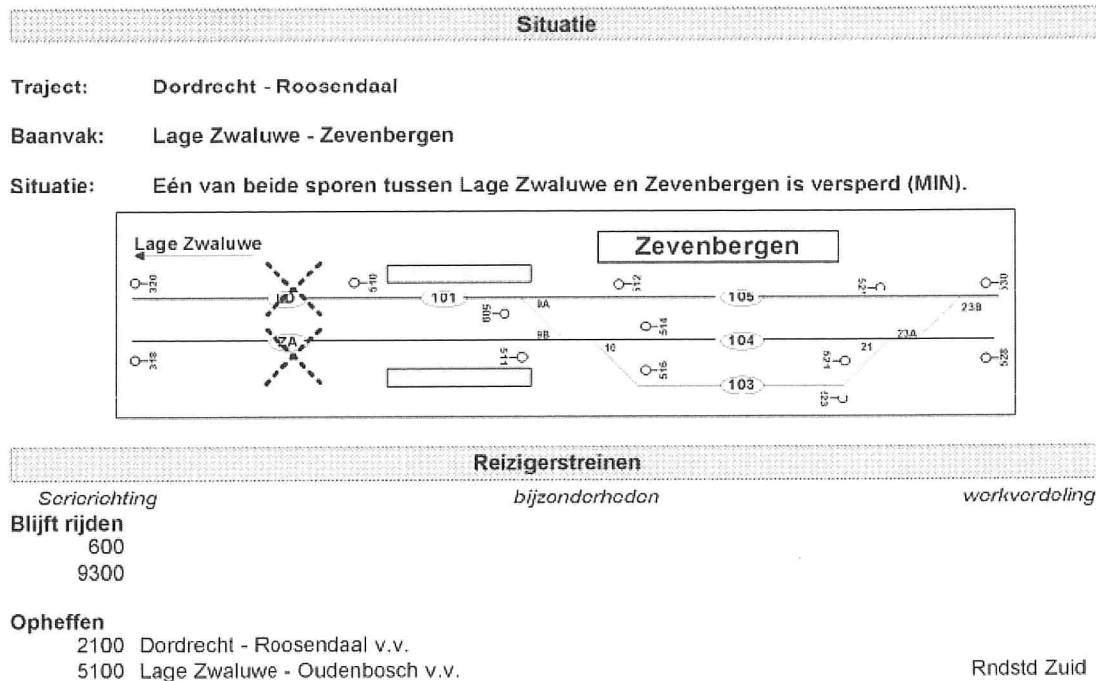


FIGURE 1 Example of emergency timetable in the Netherlands. (Source: ProRail.)

This paper presents an assessment of disruption resolution scenarios inspired by the concept of emergency timetables. The goal is to create railway plans able to recover seriously disturbed traffic conditions in large networks. The optimization-based train dispatching support system known as ROMA (railway optimization by means of alternative graphs) is applied for managing railway traffic flows on a large network characterized by a serious shortage of infrastructure capacity. Multiple performance indicators are investigated to evaluate thoroughly the alternative plans and support the dispatching process. Because of the inherent problem's complexity, the network is divided into local areas, each controlled by a local dispatcher, and a network coordinator sets constraints at the border between areas.

The next sections of this paper are organized as follows: first, the railway terminology will be introduced; second, the literature on disturbance management on large networks will be discussed; third, the framework to handle disrupted traffic situations will be given, together with a brief description of the mathematical models; fourth, computational experiments on a real railway network in the Netherlands will be provided to simulate the application of the framework in a laboratory environment; finally, the paper's conclusions will be summarized and further research directions will be outlined.

### RAILWAY TERMINOLOGY AND DEFINITIONS

Safety rules between trains dictate that at most one train at a time can occupy a block section, that is, a track segment between two main signals. In accordance with such rules, the sequence of trains can be

modified to limit delay propagation and avoid any deadlock situation. A set of trains causes a deadlock (circular waiting condition) in which each train in the set claims a block section ahead that is not available, as a result of either a disruption or the occupation of or reservation for another train in the set (4).

A microscopic timetable describes the movement of all trains running in the network during a given hour, specifying planned arrival and passing times at a set of relevant points along the route of trains (e.g., stations, junctions, and exit point of the network). The microscopic formulation refers to explicitly considering individual block sections, that is, a track segment between two main signals that may host at most one train at a time. The passage of a train on a block section is called an operation. The route of a train is a sequence of operations to be performed in a dispatching area during a service. Each operation requires a given running time, which depends on the actual speed profile followed by the train while traversing the block section.

The minimum time separations between the trains translate into a minimum setup time (time headway) between the exit of a train from a block section and the entrance of the subsequent train into the same block section. In this paper's terminology, a conflict occurs when two or more trains claim the same block section simultaneously and a decision on the train ordering has to be made that will result in changing the running time according to the speed constraints of the signaling system for at least one train. At stations, a train is not allowed to depart from a platform stop before its scheduled departure time and is considered late if it arrives after its scheduled arrival time. The resulting train delays are computed as follows. The total delay is the differ-



ence between the calculated train arrival time and the scheduled time at a relevant point in the network and is divided into two parts. The initial delay is caused by disturbances (e.g., failures, entrance delays, blocked tracks) and cannot be recovered by rescheduling train movements, except by exploiting available time reserves, that is, running trains at a maximum speed profile. The consecutive delay (knock-on delay) is caused by the interaction between trains running in the network during a given time horizon of traffic prediction.

Disruptions are heavy reductions to available capacity, caused by a train malfunction, infrastructure failures, adverse weather, or unavailability of block sections of a track. This capacity drop lasts for a long period of time required to restore the infrastructure availability, and that period can range from a few hours up to days. Under those conditions, the time reserves in the timetable are not sufficient to prevent delay propagation.

## LITERATURE REVIEW

The task of traffic controllers during disrupted operations encompasses several problems that are quite interdependent, namely, the problems of crew and rolling stock scheduling and the scheduling of train movements. With regard to the former, Jespersen-Groth et al. present a detailed report on models and procedures used to manage disruptions in Denmark and the Netherlands (5).

The main focus here is the complementary problem of the real-time computation of feasible emergency timetables for a situation of perturbed operations and reduced railway capacity. Among the contributions in this field, is a centralized train dispatching algorithm with a train rescheduling pattern language processing system proposed by Hirai et al. (6). In the case of severe traffic disruptions caused by accidents that may require the suspension of some train line, the algorithm is helpful for preparing practical rescheduling plans. On the basis of actual train schedules of a Japanese railway network, the authors report that the approach works satisfactorily even if only local modifications are applied to the original schedule.

Törnquist and Persson introduce a model for dispatching trains in a railway network with several merging and crossing points (7). A mixed integer linear programming problem is formulated and heuristic scheduling strategies are proposed to reduce the search space by restricting reordering and local rerouting actions. Experiments are presented for various disturbance settings on a Swedish railway network with 253 track segments traversed by up to 80 trains for a 90-min traffic prediction horizon.

An exact method has been proposed by D'Ariano et al. for network scheduling problems with fixed train routes (8). The computational experiments carried out on a Dutch railway bottleneck for multiple delayed trains show that optimal or near-optimal solutions are found in a very short computation time.

In large and busy networks, distributed rescheduling approaches (i.e., considering areas that negotiate orders and passing times) might be used to find feasible train schedules within acceptable computation times, even though the feasibility of global solutions is difficult to prove because of the myopic views of local solvers. Among the few papers on this subject, Jia and Zhang present a first approach based on fuzzy decision making for distributed railway traffic control (9). A multilevel decision process is described that consists of several regional decision centers to be coordinated. The test case is a main Chinese network with 12 stations and 12 trains.

Lee and Gosh report on a decentralized train scheduling algorithm for managing large networks (10). Analysis of a simulated

network from an eastern U.S. railroad shows that the proposed approach is stable when compared with input traffic rate perturbations of finite durations, depending on the capacity use, and unstable under permanent track blockage and communications link failures.

It is observed that most of the existing approaches do not provide extensive computational assessments and limit their analysis to simple and not very busy networks, rely on simplified models, and do not capture entirely the consequences of delays and other disruptions. In the current traffic control practice, dispatchers have a limited view of the possible actions to undertake, and decision support systems (DSSs) would be required that are able to compute dispatching solutions for large networks and serious disruptions at the signal control level. Such a microscopic detail is required to detect and solve deadlocks, and the risk of deadlock in disrupted situations is relevant and should not be ignored. That is the main reason for the research proposed in this paper.

This work builds on recent approaches that address the coordination problem for a complex and busy Dutch railway network divided into two dispatching areas and compares distributed and centralized systems (11). In Corman et al. the rescheduling algorithms are applied for a preliminary set of experiments on the disruption handling problem (12). During the dispatching process, multiple performance indicators are to be considered to address the various needs of the different stakeholders, and microscopic models are also needed for a detailed evaluation of the alternative rescheduling solutions.

## DISRUPTION HANDLING FRAMEWORK

Disruption handling is the management of railway traffic during operations to react to severe disturbances that make the timetable infeasible. The most used modifications of the timetable include canceling train services over the whole line or part of it (in this latter case, trains are short-turned in the vicinity of a disruption and directed toward a shunting area); rerouting train traffic in the disrupted dispatching area (using a locally available undisrupted track) or in other areas while still keeping the same origin and destination (when there is at least a possibility to bypass a disruption via other train routes); and changing the services provided by, for example, having trains carry out shorter services and keep rolling stock circulation and crew as independent as possible between areas.

Figure 2 presents a disruption handling procedure that considers a given set of disruption resolution scenarios as input to provide a new plan of operations.

For each disruption resolution scenario, it is assumed that an associated feasible rolling stock and crew management plan is available. In other words, the feasibility of operations will be recovered by a sequential approach with respect to the rolling stock, crew, and train services. The first-level decision is the selection of a set of disruption resolution scenarios, and the second-level decision is the implementation of a chosen scenario. For the latter decision, a new microscopic plan of operations is computed such that all trains run safely in the network and deviations from the timetable are minimized.

The network is operationally managed by the large-scale dispatching DSS divided into local areas of limited size, according to the general setup of Figure 2. The lower part of the scheme refers to actual train operations on the tracks. A time–distance graph is used to visualize a feasible schedule for train operations in each dispatching area. The central part of the scheme indicates dispatchers that

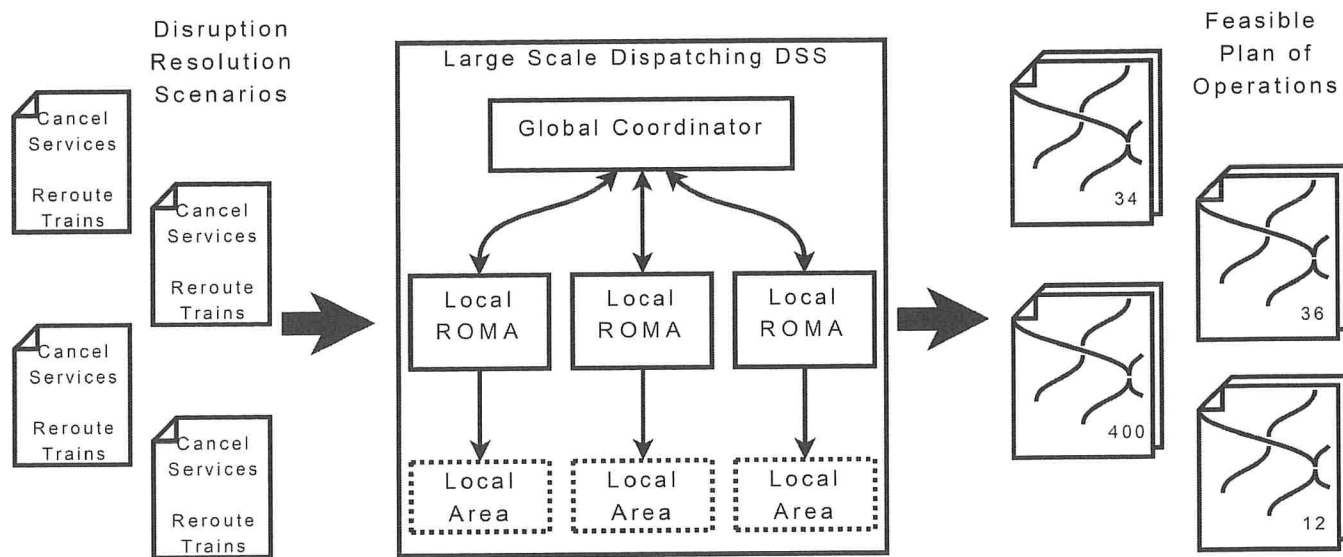


FIGURE 2 Scheme of disruption handling approach.

take actions at the level of each dispatching area. Their task is supported by local ROMA schedulers that schedule train movements according to their actual position and speed. The upper part of the scheme reports the coordination level. A global coordinator is in charge of ensuring feasibility for the overall network. This task is supported by a global coordinator module that supervises the lower levels and makes global rescheduling decisions. The final output of the DSS consists of a set of feasible operations plans, among which the dispatchers would choose the disruption resolution scenario for implementation.

### MODEL OF LOCAL ROMA SCHEDULERS

The approach used to reschedule trains in each local dispatching area is a microscopic problem formulation based on the alternative graph (13), a job shop scheduling formulation with additional constraints, and the blocking time theory (14), a recognized method to compute time separations between operations.

To solve the dispatching problem on each area, the branch and bound algorithm described by D'Ariano et al. is used (8). The exact algorithm is based on the alternative graph and computes an optimal train schedule that minimizes the maximum consecutive delay. This algorithm uses various speedups based on the infrastructure topology and algorithmic improvements based on graph properties.

The alternative graph has been used to model and solve train scheduling problems in several papers for its detailed and flexible representation of a network (8, 12). Each operation corresponds to a node in the alternative graph, and the arcs between nodes are used to model the blocking times. The alternative graph represents the routes of all trains in a given control area; because a train must traverse the block sections in its route sequentially, a train route is modeled in the alternative graph with a job that is a chain of operations and associated precedence constraints. A train schedule corresponds to the set of the starting time of each operation. A potential conflict between two trains on a block section is modeled as a pair of alternative arcs for each pair of trains traversing the block section. A deadlock-free and conflict-free

schedule is obtained in such a way that there is no positive length cycle in the graph.

### MODEL OF GLOBAL COORDINATOR

To solve the network coordination problem, an approach is used that decomposes the global scheduling problem of scheduling dense train traffic in a large area into smaller affordable subproblems, which can be solved independently of each other (11). The local solutions are then coordinated to obtain a global solution to the original problem.

The coordination level makes use of an aggregated model introduced by Corman et al. for the rescheduling decisions computed for each local area (11). The model is based on a compact representation of the information concerning the local solutions, such as a minimum temporal distance between the entering and leaving of a train in each local area and the entrance and exit times of each train traversing area borders. This information is used to build a border graph, useful to check the global feasibility of the local solutions found and to guide a heuristic coordination procedure to iteratively harmonize the solutions found by local solvers (11). This results in imposing coordination constraints on the local schedulers, such as precedence constraints between trains at border nodes or a particular value for the exit time from an area and the entrance time in the subsequent area for a train. Iteratively, each local scheduler computes a new schedule with such additional constraints until a globally feasible solution is found, a local infeasibility is found, or a computation time limit is reached.

### TEST CASE

A real-world test case is considered that is based on a large part of the Dutch railway network (Figure 3), including the main stations of Den Bosch, Nijmegen, Arnhem, and Utrecht.

Figure 3 shows the network divided into three local dispatching areas. The general network layout is a circular shape of about 300 km in length with more than 1,200 block sections and station stopping

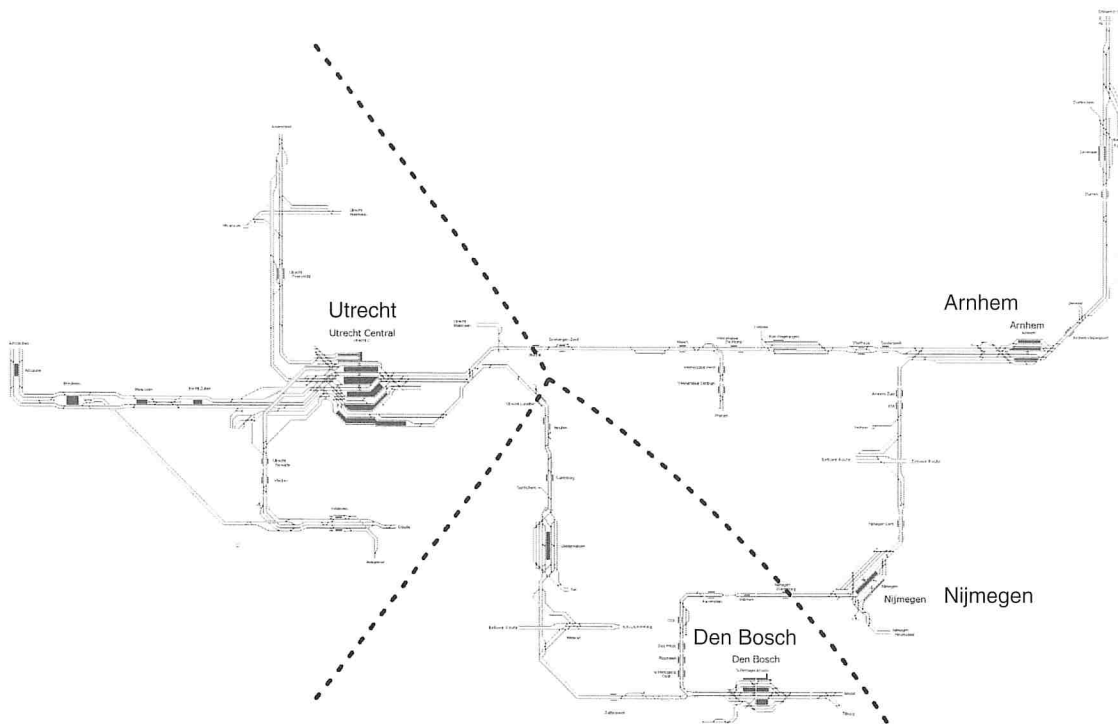


FIGURE 3 Railway network and its division into three local areas.

platforms. The reference timetable is periodic and schedules about 150 trains per hour.

### INFRASTRUCTURE DISRUPTION AND RESOLUTION SCENARIOS

An infrastructure disruption located on the Utrecht–Den Bosch line (near Zaltbommel) is studied; it blocks one track and reduces the maximum train speed allowed on the adjacent track (60 km/h instead of 130 km/h). In Figure 4 the track blockage is outlined by the red star along the line between Utrecht and Den Bosch. In the timetable, 12 trains per hour (six per direction) are scheduled on the disrupted line.

Managing this disrupted situation may require canceling services or short-turning trains at the major stations of Utrecht and Den Bosch or local rerouting to avoid the disrupted track. In the latter case, trains of both traffic directions have to run on the available single track and under additional constraints on the maximum speed allowed for a railway stretch about 6 km long. Another possibility is to reroute trains globally along a different line that goes from Den Bosch to Nijmegen and Arnhem to Utrecht, and vice versa. In this additional case, the travel time for the alternative trip between Utrecht and Den Bosch is about 40 min longer than the original trip time, which is 30 min long.

In Figure 4 the scenarios for the resolution of disruptions are identified by the three field code *a-b-c*; *a* is the number of trains locally rerouted to bypass the disruption, *b* is the number of trains globally rerouted via Arnhem and Nijmegen, and *c* is the number of canceled services. Those disruption resolution scenarios have been derived by incrementally canceling trains, rerouting trains, or following shuttle operations.

The following scenarios are based on the original timetable:

- 12-0-0. All trains are scheduled as in the reference timetable, locally rerouted in the vicinity of the disruption along the only available track.
- 8-4-0. Four intercity trains and four local trains are still scheduled on the Utrecht–Den Bosch line. The other four intercity trains are globally rerouted via Nijmegen and Arnhem, that is, their routes are changed but the same origin and destination stations are kept.
- 8-0-4. Four intercity and four local trains are scheduled on the Utrecht–Den Bosch line, and the other four intercity services are canceled, which results in fewer trains running in the network. The corresponding train units are thus held in the major stations of Utrecht and Den Bosch, if there is enough spare capacity at the stations.
- 4-4-4. Four local trains are still scheduled on the Utrecht–Den Bosch line, four intercity trains are globally rerouted via the Nijmegen–Arnhem line, and four intercity trains are held in the major stations of Utrecht and Den Bosch.
- 4-0-8. Four local trains are still scheduled on the Utrecht–Den Bosch line, and eight intercity trains are held in the major stations of Utrecht and Den Bosch. The local trains thus serve all passengers.

An alternative proposal for the design of emergency timetables is based on shuttle services for which trains arrive at the designated station and leave from it to operate a new service in the opposite direction after a turnaround time. Shuttle timetables have been recently considered by Dutch railway managers, particularly to handle the risk of countrywide delays that would affect the rolling stock and crew circulation (e.g., in the case of heavy snowfall). In the shuttle timetable, (*a*) train paths are defined as short as possible, so that the propagation of delays cannot exceed regional boundaries;

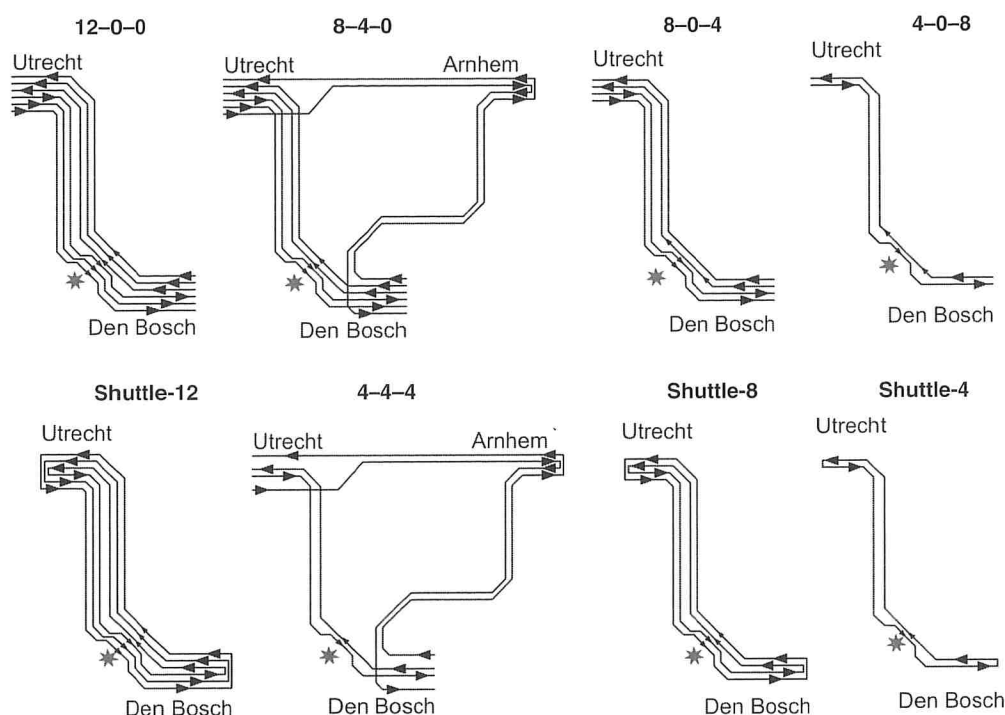


FIGURE 4 Disrupted traffic situation and eight resolution scenarios.

(b) passenger connections might be kept for each trip interrupted by the disruption; and (c) the circulation plan of drivers and rolling stock units is combined to avoid the inconvenience of trip repositioning, which may cause further knock-on effects between train lines.

Next, shuttle timetables are combined with cancellation of services to limit the additional rolling stock and crew units required. It is assumed that each crew and rolling stock unit has the same circulation and that the shuttle services exploit local rerouting to avoid track blockage. The following timetables described from the point of view of the corridor between Utrecht and Den Bosch are considered:

- **Shuttle-12.** The frequency of services is kept as in the original timetable. Eight intercity services, evenly distributed along the timetable hour, provide services to travelers, together with four local trains.
- **Shuttle-8.** Four intercity services plus four local trains serve all passengers, the other four intercity trains having been canceled.
- **Shuttle-4.** Four local trains serve all passengers.

The microscopic formulation of the shuttle services requires the detailed modeling of shunting movements in interlocking areas. Also required is the introduction of connection constraints (as in D'Ariano et al.) between the arrival of train unit and crew and the departure of the next service (15).

## DISPATCHER SUPPORT INTERFACE

Figure 5 reports a blocking time diagram for Scenario 8-4-0 in the vicinity of the disruption (represented by the red cross between the locations labeled Ht and Zbm, in the top part of the figure) as a support interface for dispatchers. The first 45 min of traffic prediction

is reported on the y-axis because these are most relevant in understanding the short-term effects of delay propagation. The train traffic on the Utrecht–Den Bosch line is shown on the x-axis (the line is about 40 km long) of Figure 5. Specifically, the following stations are considered, from left to right: Den Bosch (Ht), Zaltbommel (Zbm), Geldermalsen (Gdm), Culemborg (Cl), Houten (Htn), and Utrecht Lunetten (Utl). The main traffic goes from top left to bottom right, but there are also trains going from top right to bottom left (highlighted in gray). The latter trains need to run against their normal travel direction on the only available track.

When solving disrupted traffic situations, the dispatchers can easily detect the delayed trains by observing the long stretched blocking times. In the example in Figure 5, trains wait in front of a red signal to enter the single track area. The available spare capacity can be measured as the time lag between consecutive blocking times. When there is no spare capacity between the passage of two or more trains on the same block section (i.e., the corresponding time lag is null), the trains following may increase their running times on the previous block sections to avoid a conflict with the other trains.

## PERFORMANCE INDICATORS

The following performance indicators are considered:

- “Generalized travel time” for a given origin–destination (O-D) pair is defined as the weighted sum of waiting times at the origin station (weighted 1.5) and at the intermediate stations (weighted 4) plus the travel time between stations (weighted 1). As in Wardman, it is assumed that passengers penalize a long waiting time at a station more than a long travel time (16).
- “Minimum cycle time” for a given O-D pair is the cycle time needed to dispatch all timetable services. For the hourly timetable, if

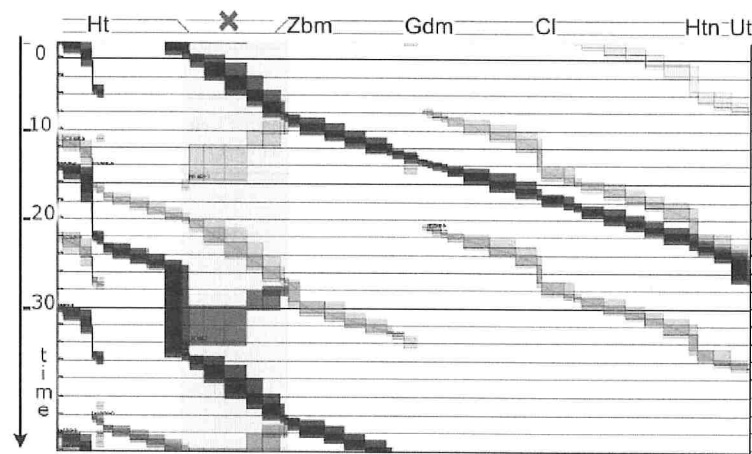


FIGURE 5 Blocking times for Scenario 8-4-0, Utrecht–Den Bosch dispatching area.

this time is longer than 1 h, the next timetable hour will be affected by delay propagation.

- Service frequency is expressed as the number of services for a given O-D pair in a given time. This is a key factor in forecasting passenger flows and avoiding overcrowded trains.

- “Punctuality” is the percentage of trains that are running with at most 3 min of total delay. This is a standard performance indicator to evaluate the quality of operations in the Netherlands. Train delays are also considered in regard to total delays and consecutive delays.

- Variations of the required rolling stock (and associated crew) units related to shuttle or canceled services are considered in the additional rolling stock units, when compared with the original timetable.

## ASSESSMENT OF SCENARIOS

For each disruption resolution scenario, train schedules are computed for 90-min time horizons of traffic prediction. The disruption is combined with a set of 30 delay instances (maximum entrance delay is about 900 s and average entrance delay is about 180 s) to study the combined effect of perturbances. For the evaluation of the various scenarios, the dispatching system introduced previously was run on

an Intel iCore 5 2.66 Ghz computer with 4 GB RAM in an average time of 3 min per scenario.

Table 1 presents a set of performance indicators related to the train delays for each evaluated scenario. Each row of this table presents the average results over the 30 delay instances. Column 1 reports the reference scenario, Columns 2 and 3 the maximum and average total delays (in seconds), Column 4 the percentage of punctual trains, Columns 5 and 6 the maximum and average consecutive delays (in seconds), and Column 7 the variation (positive or negative) in the number of rolling stock (and associated crew) units required to run all services. The best value of each column is emphasized in bold.

The results in Table 1 show that Scenario 4-0-8 is the best for most of the delay indicators. The other most favorable solutions are 8-0-4 and Shuttle-12, with the highest punctuality score, and Shuttle-4 with the lowest average total delay score. The better delay reduction obtained for Scenario 4-0-8 is motivated by the reduced number of running trains (eight intercity train services are canceled). The better punctuality obtained for Shuttle-12 is counterbalanced by the larger total and consecutive delays compared with Scenario 4-0-8. In other words, fewer trains are delayed but delays are greater.

Passenger point of view is considered for a limited set of relevant O-D pairs because of a lack of detailed information about

TABLE 1 Delay Indicators for Scenarios of Disrupted Situation

Disruption Resolution Scenario	Max. Total Delay (s)	Avg. Total Delay (s)	Percentage of Punctual Trains (<3 min)	Max. Cons. Delay (s)	Avg. Cons. Delay (s)	Additional Rolling Stock Units
12-0-0	2,623	383	52	1,816	160	0
8-4-0	3,656	444	47	1,024	110	0
8-0-4	1,513	223	<b>65</b>	792	60	−4
4-4-4	3,509	369	56	918	85	−4
4-0-8	<b>1,186</b>	213	64	<b>616</b>	<b>44</b>	<b>−8</b>
Shuttle-12	2,477	274	<b>65</b>	1,867	148	8
Shuttle-8	1,586	262	52	780	68	4
Shuttle-4	1,239	<b>211</b>	60	636	49	0

NOTE: Max. = maximum; avg. = average; cons. = consecutive.



TABLE 2 Other Indicators for Scenarios of Disrupted Situation

Disruption Resolution Scenario	Den Bosch to Amsterdam			Den Bosch to Utrecht		
	Service Frequency	Generalized Travel Time (s)	Min. Cycle Time (s)	Service Frequency	Generalized Travel Time (s)	Min. Cycle Time (s)
12-0-0	<b>8</b>	5,326	4,804	<b>11</b>	5,731	4,352
8-4-0	<b>8</b>	5,506	4,348	9	5,513	3,958
8-0-4	4	<b>4,933</b>	4,111	8	5,497	3,766
4-4-4	1	8,137	4,246	4	5,258	3,836
4-0-8	0	—	—	3	5,169	3,698
Shuttle-12	1	6,664	4,440	<b>11</b>	5,489	3,831
Shuttle-8	1	7,249	3,874	9	5,658	3,645
Shuttle-4	1	7,226	<b>3,754</b>	3	<b>5,164</b>	<b>3,562</b>

NOTE: — = not applicable; min. = minimum.

actual passenger flows in the network. Clearly, studying all O-D pairs would make it possible to evaluate networkwide measures of passenger discomfort more comprehensively.

On the space–time network defined by the different train services, including possible intermediate stops and associated transfer times, the shortest paths are computed for the passengers of the selected O-D pairs. In Table 2, the service frequency is reported over the given traffic prediction time horizon, generalized travel time, and minimum cycle time for the following O-D pairs: Den Bosch–Amsterdam Central Station and vice versa and Den Bosch–Utrecht Central Station and vice versa. The former two pairs may include passenger transfers in Utrecht Central Station. Each row presents the average results over the 30 delay instances for each evaluated scenario. Again, the best value of each Table 2 column is emphasized in bold.

Most of the disruption resolution scenarios have problems in limiting the propagation of train delays because the minimum cycle time is often greater than 3,600 s. This effect is more evident when the frequency of services is increased. In general, shuttle timetables offer a better minimum cycle time than do the other disruption resolution scenarios. Because of the large disturbances and related delay propagation, the train schedules computed for Scenarios 4-0-8, 4-4-4, and Shuttle-4 are not able to serve all passengers traveling on the Den Bosch–Amsterdam line within the 90-min traffic predictions. In regard to the generalized travel time, the combined effect of canceling services (see, e.g., Scenario 8-0-4 for the Den Bosch–Amsterdam direction or Scenario Shuttle-4 for the Den Bosch–Utrecht direction) and rerouting trains (see, e.g., Scenario 4-4-4 for the Utrecht–Den Bosch direction) produces good results on a specific O-D pair, but that combined effect cannot be easily generalized to all pairs.

## CONCLUSIONS AND FURTHER RESEARCH

This paper applies a state-of-the-art decision support system for railway dispatching to handle the railway disruption management process. An advanced disruption handling approach is adopted to compute feasible train schedules at the microscopic level for exceptional situations. Disruption resolution scenarios and advanced scheduling algorithms are combined to obtain feasible dispatching plans in a short computation time, even for large areas with up to thousands of block sections and hundreds of trains. On each solution, the negative effect of disruptions is quantified with respect to passenger travel and waiting times, train

delays, punctuality, timetable cycle time, and service frequency. This calculation would allow the dispatchers to choose the most effective scenario and the corresponding microscopic plan of operations.

Future research should address the implementation of advanced interfaces between human traffic controllers and DSSs, on the basis of key performance indicators for each dispatching solution and disruption resolution scenario. Multiple scenarios and dispatching solutions to be evaluated could be automatically generated on the fly, on the basis of the actual infrastructure availability. Such an approach could also be used in the planning stage to evaluate the feasibility and performance of alternative timetables. After the timetable assessment, the rolling stock and crew schedules would also have to be updated accordingly. Other applications of the dispatching support tool should be the validation of countrywide timetables and the development of integrated approaches for rolling stock and crew rescheduling.

## ACKNOWLEDGMENTS

The authors thank ProRail for providing the instances. This work is partially supported by the Italian Ministry of Education, University, and Research (MIUR), and its Project FIRB Advanced Tracking System in Intermodal Freight Transportation.

## REFERENCES

1. Clausen, J. Disruption Management in Passenger Transportation—From Air to Tracks. *Proc., 7th Workshop on Algorithmic Approaches for Transportation Modeling, Optimization, and Systems*, Seville, Spain, 2007.
2. Meng, X., L. Jia, and Y. Qin. Train Timetable Optimizing and Rescheduling Based on Improved Particle Swarm Algorithm. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 2197, Transportation Research Board of the National Academies, Washington, D.C., 2010, pp. 71–79.
3. Meng, L., and X. Zhou. Robust Train-Dispatching Model Under a Dynamic and Stochastic Environment: A Scenario-Based Rolling Horizon Solution Approach. Presented at 90th Annual Meeting of the Transportation Research Board, Washington, D.C., 2011.
4. Coffman, E. G., M. J. Elphick, and A. Shoshani. System Deadlocks. *Computing Surveys*, Vol. 3, No. 1, 1971, pp. 67–78.
5. Jespersen-Groth, J., D. Pothhoff, J. Clausen, D. Huisman, L. G. Kroon, G. Maróti, and M. N. Nielsen. Disruption Management in Passenger

Amsterdam to Den Bosch			Utrecht to Den Bosch		
Service Frequency	Generalized Travel Time (s)	Min. Cycle Time (s)	Service Frequency	Generalized Travel Time (s)	Min. Cycle Time (s)
5	5,698	4,352	14	5,616	4,352
5	6,159	3,958	14	5,706	3,958
2	6,146	3,766	9	5,595	3,766
0	—	—	4	<b>4,683</b>	3,836
0	—	—	1	6,478	3,698
<b>6</b>	<b>5,570</b>	3,831	<b>21</b>	5,839	3,831
1	5,603	<b>2,991</b>	9	5,078	3,645
0	—	—	2	5,439	<b>3,562</b>

- Railway Transportation. *Lecture Notes in Computer Science* 5868, 2009, pp. 399–421.
6. Hirai, C., N. Tomii, Y. Tashiro, S. Kondou, and A. Fujimori. An Algorithm for Train Rescheduling Using Rescheduling Pattern Description Language R. *Computers in Railways X*, 2006, pp. 551–561.
  7. Törnquist, J., and J. A. Persson. N-Tracked Railway Traffic Re-Scheduling During Disturbances. *Transportation Research Part B*, Vol. 41, No. 3, 2007, pp. 342–362.
  8. D'Ariano, A., D. Pacciarelli, and M. Pranzo. A Branch and Bound Algorithm for Scheduling Trains in a Railway Network. *European Journal of Operational Research*, Vol. 183, No. 2, 2007, pp. 643–657.
  9. Jia, L.-M., and X.-D. Zhang. Distributed Intelligent Railway Traffic Control: A Fuzzy-Decisionmaking-Based Approach. *EAAI*, Vol. 7, No. 3, 1994, pp. 311–319.
  10. Lee, T. S., and S. Gosh. Stability of RYNSORD: A Decentralized Algorithm for Railway Networks Under Perturbations. *IEEE VT*, Vol. 50, No. 1, 2001, pp. 287–301.
  11. Corman, F., A. D'Ariano, D. Pacciarelli, and M. Pranzo. Centralized Versus Distributed Systems to Reschedule Trains in Two Dispatching Areas. *Public Transport*, Vol. 2, No. 3, 2010, pp. 219–247.
  12. Corman, F., A. D'Ariano, and I. A. Hansen. Disruption Handling in Large Railway Networks. *Computers in Railways XII*, 2010, pp. 629–640.
  13. Mascis, A., and D. Pacciarelli. Job Shop Scheduling with Blocking and No-Wait Constraints. *European Journal of Operational Research*, Vol. 143, No. 3, 2002, pp. 498–517.
  14. Hansen, I. A., and J. Pahl. *Railway Timetable and Traffic: Analysis, Modelling and Simulation*. Eurailpress, Hamburg, Germany, 2008.
  15. D'Ariano, A., F. Corman, D. Pacciarelli, and M. Pranzo. Reordering and Local Rerouting Strategies to Manage Train Traffic in Real-Time. *Transportation Science*, Vol. 42, No. 4, 2008, pp. 405–419.
  16. Wardman, M. Public Transport Values of Time. *Transport Policy*, Vol. 11, No. 4, 2004, pp. 363–377.

*The Intercity Passenger Rail Committee peer-reviewed this paper.*

# High-Speed Rail Versus Air Transportation

## Case Study of Madrid–Barcelona, Spain

Francesca Pagliara, José Manuel Vassallo, and Concepción Román

Travel time savings, better quality of supplied services, greater comfort for users, and improved accessibility are the main factors of success for high-speed rail (HSR) links. In this paper, results are presented from a revealed and stated preference survey concerning HSR and air transport users in the Madrid–Barcelona, Spain, corridor. The data gathered from the stated preference survey were used to calibrate a modal choice model aimed at explaining competition between HSR and air transportation in the corridor. The major findings of the paper describe the demand response to different policy scenarios considering improvements in the level of transport services. From the model, prices and service frequency were found to be among the most important variables in competing with the other mode. In addition, it was found that check-in and security controls at the airport are a crucial variable for users in making their modal choices. Other policies, such as the improvement of parking facilities at train stations, play a secondary role.

The development of high-speed rail (HSR) has been one of the central features of the recent European Union (EU) transport infrastructure policy. The proposals for a European HSR network emerged in a report of the Community of European Railways in 1990, and this was adopted as the base for what essentially became the European Community's proposed Trans-European Network (1). The latter, which is basically the linking of a series of national plans for promoting HSR improvements, emerged during the 1970s and 1980s.

High-speed trains can be used to solve two different accessibility problems. In the first case, in which a point-to-point link is dominant, each train is a potential substitute for an air connection between two cities (2, 3). The HSR link Madrid–Barcelona, Spain, belongs to this case. In the second case, in which a high-speed network is dominant, the train system links many cities and central business districts and, therefore, creates a new type of region sharing a common labor market and a common market for household and business services.

Having first sanctioned a 160-km/h maximum speed as recently as 1986, Spain moved quickly to get HSR into operation. The first prestigious *alta velocidad Española* (AVE), used to denote long-distance HSR services, linked Madrid with the country's fourth largest city in Spain, Seville, which had been chosen to host the 1992 Expo World's

Fair and to stimulate the economy of the country's south in general. It was not until 2008 that Barcelona would have gained the AVE link with Madrid.

HSR was so successful in Spain that in its latest National Infrastructure Plan the government decreed that all capitals of Spain's provinces should have a high-speed connection no longer than 4 h from the capital, Madrid, and 6.5 h from the second city, Barcelona (4). However, because of the high infrastructure cost of HSR and the shortfall of budgetary resources caused by the economic recession starting in 2008, the government has since postponed or even canceled some of the projects already approved, such as the connection with Portugal.

The objective of this paper is to identify the key aspects that explain mode choice between HSR and frequent air transportation services for HSR in the Madrid–Barcelona corridor. The methodology used is based on modeling the choice between air and rail through the calibration of a binomial logit model with a survey carried out between February and March 2010.

The effects caused by investments in HSR have been analyzed in the literature in many different ways. In particular, studies on the Madrid–Barcelona corridor carried out before the entrance of HSR can be classified into the following groups: (a) evaluations of the economic profitability of particular corridors or areas [see de Rus and Román for the Madrid–Zaragoza–Barcelona HSR (5), de Rus and Nombela for the EU (6), and Martín and Nombela for Spain (7)]; (b) studies of the effects on accessibility (8–10); and (c) studies on intermodal competition based on the analysis of passenger perceptions and preferences (11, 12), which analyzed the potential of high-speed trains to compete with airlines and private car markets by using stated preference (SP) experiments.

López-Pita and Robusté analyzed the effect that high-speed railway services have on air traffic demand by using forecasting models that have been applied in Europe (13). Forecasts predicted that the railway line will have a market share of between 53% and 63%, compared with its current 11%, thus reducing the airlines' current 89% market share to between 36% and 47%. More recently, Román et al. used a mixed revealed preference (RP)–SP data set to study the Madrid–Zaragoza–Barcelona line, focusing on modeling issues and policy analysis (14, 15). Effort was concentrated on Madrid–Zaragoza and Madrid–Barcelona routes, in which HSR could attract more traffic from the competing modes. Román and Martín predicted an expected demand for HSR in the Madrid–Barcelona corridor of between 2.7 million and 3.2 million passengers per year, with a market share for HSR in the air–rail market ranging from 43% to 48%, and pointed out that this volume of traffic is not enough to guarantee that this project will have a positive social benefit (16). Finally, Román and Martín highlighted the important role that access time to terminals may play in regard to modal competition between rail and plane for interurban travel passengers (17).

F. Pagliara, Department of Transportation Engineering, University of Naples Federico II, Via Claudio 21, Naples 80125, Italy. J. M. Vassallo, Departamento de Transportes, ETSI de Caminos, Canales y Puertos, Universidad Politécnica de Madrid, Profesor Aranguren s/n, Madrid 28040, Spain. C. Román, Facultad de Economía, Empresa y Turismo, University of Las Palmas de Gran Canaria, Modulo D, Campus de Tafiara, Las Palmas de Grand Canary 35011, Spain. Corresponding author: J. M. Vassallo, jvassallo@caminos.upm.es.

*Transportation Research Record: Journal of the Transportation Research Board*, No. 2289, Transportation Research Board of the National Academies, Washington, D.C., 2012, pp. 10–17.  
DOI: 10.3141/2289-02

This paper is organized as follows. The next section describes the main features of competition between HSR and air transportation in the Madrid–Barcelona corridor. In the following section, a mode choice model is calibrated with data from a survey carried out after the introduction of HSR. The last section discusses the conclusions and further perspectives.

## COMPETITION BETWEEN HSR AND AIR TRANSPORT IN MADRID–BARCELONA CORRIDOR

### Madrid–Barcelona Corridor Before HSR

Before the entrance of HSR, the Madrid–Barcelona rail line was served by a low-quality service, especially in regard to the commercial speed achieved between the two cities (13). A conventional train of the Talgo technology covered the 625-km distance in 5 h 30 min for an average ticket price of €65 and with a service frequency of eight departures per day. Before the opening of the HSR service the patronage of the former rail service was about 800,000 passengers a year.

Most of the trips between Madrid and Barcelona (about 4.8 million in 2007), were therefore made by air transportation. In fact, the Madrid–Barcelona route was the busiest air route in Europe before inauguration of the HSR system. Iberia has been the major carrier on this route, particularly because of its air shuttle service “Puente Aéreo” (PA), which moved 3 million passengers before the opening of the HSR service. This air shuttle was conceived as a commuter service. Passengers do not need a previous booking; they just arrive at the airport and board the next available flight. If a plane is full, another one departs shortly after, with peak-hour frequencies of departures every 15 min, rivaling those of public transportation. The idea is to provide plenty of flexibility and short waiting times at the airport. PA has its own identity brand and fare structure. PA has been a sort of second home to generations of businessmen and politicians, ready to pay for all this flexibility and convenience. It has been for years Iberia’s most profitable route; in a way, PA is to Iberia what London Heathrow–New York is to British Airways.

After the liberalization of the Single European Sky, other air carriers such as Spanair, Air Europa, and Vueling entered the Madrid–

Barcelona market. Even though these companies have been gaining share over the years, they have not been able to beat the hegemony of Iberia on this route.

### Opening of HSR Service Between Madrid and Barcelona

The Madrid–Barcelona corridor is one of the busiest in passenger transport in Europe. Madrid City has a population of about 3.3 million inhabitants, but the Madrid metropolitan area has about 6.5 million. The city of Barcelona has 1.6 million, but its metropolitan area reaches 3 million. Both cities are quite compact. The density of Madrid is 5,400 inhabitants per square kilometer, and the density of the city of Barcelona is 15,900 inhabitants per square kilometer, more compact than the density of most U.S. cities.

The completion of HSR between Barcelona and Madrid has had a stronger effect on the route as HSR has emerged as a real alternative in regard to frequency and comfort for business travelers, taking half of the market on the Madrid–Barcelona corridor. HSR has forced Iberia to reduce capacity and maintain frequency with smaller aircraft. However, the 625-km (388-mi) distance between Madrid and Barcelona is really at the edge of what is considered a competitive distance range for HSR. For that reason the analysis of competition between HSR and PA in this corridor was found to be particularly interesting.

Figure 1 shows the evolution of market share before and after introduction of the new HSR. The total number of passengers traveling by air and train has increased constantly over the years, reaching a peak in 2008. The reduction of passengers from 2008 on was caused by the economic recession that struck Spain and has continued ever since. Some of those passengers moved to cheaper and slower transport modes such as coaches.

With the introduction of HSR at the beginning of 2008, passenger volume for the rail mode increased by 1,380,000 in the first year of operation and by more than 500,000 in the second year. In the same period air transport lost 800,000 passengers in the first year and more than 1 million in the second year. The modal share at the end of 2009 was 47.1% for rail and 52.9% for air.

In 2010 a slight increase in modal share for the air mode was registered (to 54.4%) and a slight drop to 45.6% for rail. The reason

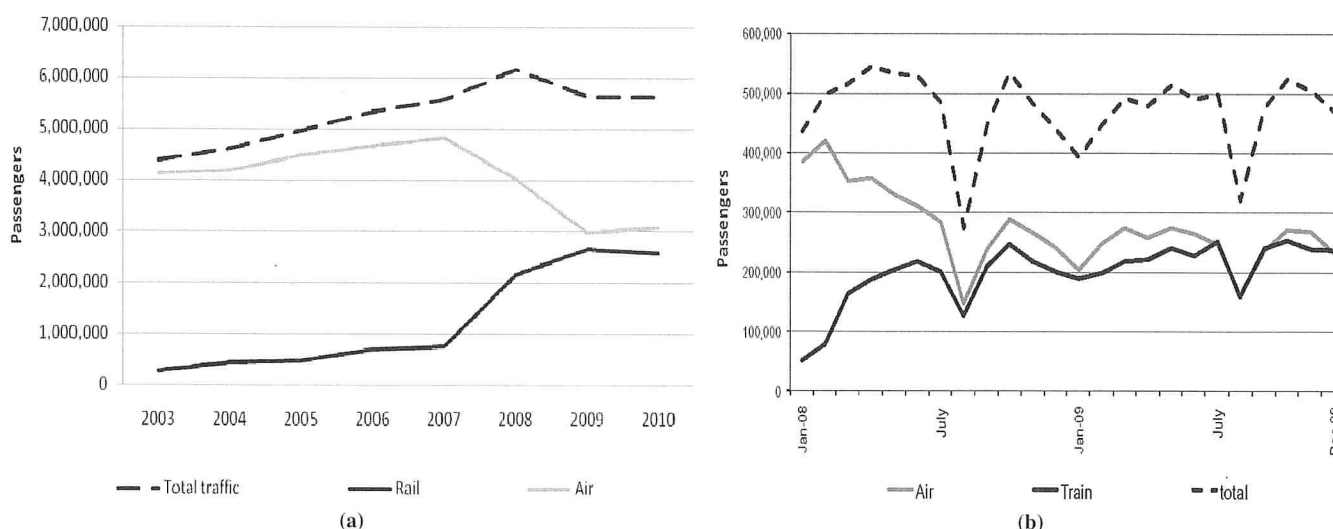


FIGURE 1 Trend of passengers along Madrid–Barcelona corridor.

for this increase was the air companies' response to the competition from HSR: the reduction of air ticket prices.

### Competitive Characteristics of HSR and Iberia Air Shuttle in Madrid-Barcelona Corridor

Madrid and Barcelona are the two most important economic poles in Spain. This fact generates a great number of short-term business trips between the two cities. Before the opening of the HSR service, most of these trips were channeled through the Iberia air shuttle, PA. After the inauguration of HSR a strong competition started between PA and HSR services. This subsection analyzes the aspects that might influence the modal choice. The following have been identified: travel time, service frequency, price, reliability, and comfort, along with the possibility of engaging in additional activities during the trip.

#### Travel Time

Travel time is made up of (a) access time from origin to airport or station, (b) check-in time and security procedures, (c) waiting time at the airport or station, (d) transport time, (e) baggage pickup time if necessary, and (f) egress time from airport or station to final destination.

PA transport time is 1 h 10 min. HSR transport time depends on the number of stops made by the train. There are basically three different types of HSR service in regard to time: (a) nonstop trains that take 2 h 38 min, (b) trains stopping in Zaragoza that take 2 h 52 min, and (c) trains stopping three times along the way that take 3 h 18 min. Consequently the travel time varies between 2 h 38 min, and 3 h 18 min.

The time in the airport for the air shuttle is only 30 min. This time includes check-in, security procedures, and waiting time. This time is shorter than the equivalent for conventional flights because PA has a specific security checkpoint next to a specific boarding gate assigned to PA, which is the boarding gate closest to the airport entrance. The check-in and waiting time for HSR is between 10 and 15 min.

The baggage check-in and pickup time is nonexistent for HSR since passengers are allowed to board trains with heavy luggage. Check-in and pickup time plus security procedures time are much longer for PA. However, 80% of PA users do not check baggage, so ultimately that is not such a critical aspect. The EU sets up common rules across its countries to protect civil aviation against acts

of unlawful interference. The regulation's provisions apply to all airports or parts of airports located in an EU country that are not used exclusively for military purposes. Airport security in Spain is provided by police forces, as well as private security guards. Security checkpoints in Spain involve inconvenience for users stemming from the long time spent waiting in lines and the requirement of passing through detectors for metal and explosives.

Access time from the origin of the trip to the departure airport or station and the egress time from the arrival airport or station to the final destination are crucial factors explaining modal choice. Atocha Station in Madrid and Sants Station in Barcelona are located in central positions inside the cities with good accessibility by public transportation. However, the stations are not convenient for people getting there by car because parking facilities in the stations are small and fill up early in the morning. Barajas Airport in Madrid and El Prat Airport in Barcelona are located 15 km and 13.5 km away from their respective city centers. This is not too far compared with other airports. Despite that, Barajas and El Prat Airports are still much farther from their respective city centers than are Atocha and Sants Stations.

Table 1 shows that average travel times by public transportation to the stations are shorter than to the airports, especially in Barcelona. This table displays average travel times. Obviously personal travel times depend on the ultimate location of the origin and the destination of the trip within the regions. The difference in regard to car accessibility, by both private car and taxi, is not so notable. Moreover, even though travel times by car and taxi to the stations are shorter, Atocha and Sants Stations have two problems. The first one is their scarce parking capacity. Atocha Station in Madrid has only 965 parking spaces compared with 16,300 spaces at Barajas Airport, and Sants Station Barcelona has only 900 parking spaces compared with 13,000 at El Prat Airport. The second one is waiting time for passengers taking taxis, which is longer at the stations compared with waiting time at the airports.

#### Travel Cost

HSR and PA have different travel classes. PA has the traditional economy and business classes, even though most of the passengers (98% according to the questionnaire conducted for this research) choose economy class. The HSR service has three different classes: economy class, which is chosen by most passengers (81% according to the questionnaire conducted for this research); business class chosen by

TABLE 1 Average Access Time and Average Cost to Get to HSR Stations and Airports in Madrid and Barcelona

Access Mode	Madrid				Barcelona			
	Time (min)		Cost (€)		Time (min)		Cost (€)	
	Atocha Station	Barajas Airport	Atocha Station	Barajas Airport	Sants Station	El Prat Airport	Sants Station	El Prat Airport
Taxi	15	20	14	28	10	25	10	26
Car	15	20	29	20	10	25	26	20
Local train	15	NA	0	NA	12	45	0	3
Metro	30	35	1	2	12	NA	1	NA
Bus	45	45	1	1	20	40	1	5

NOTE: NA = not available.



18%; and club class chosen by 1%. In this section the focus is on cost for economy class, which is the service taken by most users.

Travel cost includes the price of buying the air or train ticket plus the access cost to and from the stations. Ever since inauguration of the HSR service, there have been constant changes in the pricing policies of the PA and HSR services responding to each other's strategy. To give an example, when HSR service between Madrid and Barcelona was inaugurated, PA reduced its prices by 35%, the lowest air fare for this service.

In March 2010, when the survey described in the following section was conducted, the one-way price for PA was €129. Iberia offered rebates to this price for either buying a round-trip ticket or buying the ticket in advance. The one-way price for HSR was €135.5, slightly more expensive than PA's price. RENFE, the company operating the HSR service between Madrid and Barcelona, offered a 20% discount for a round-trip ticket, a 40% discount for buying the ticket 1 week in advance, and a 60% discount for buying the ticket 2 weeks in advance.

To the price of the ticket, the cost of accessing and leaving the airport or station has to be added. Table 1 shows the average costs to get to and to come from the stations and airports in Madrid and Barcelona. The costs include parking fares paid in the case of parking a car at the airport or station. The cost of using a local train to get to or to leave the stations is zero because the HSR ticket enables users to take a local train at the station free of charge. The only access mode that is more expensive for getting to the stations than to the airports is the car, precisely because of the high price of parking at Madrid and Barcelona stations compared with parking prices at the airports.

### *Service Frequency*

Frequency is a crucial variable for modal choice, particularly for business travelers. More frequent services allow passengers greater flexibility to get back sooner or later if their meetings end before or after schedule. That issue explains why, despite the PA patronage reduction after the opening of the HSR service, one of Iberia's priorities for PA was to maintain the same frequency. This was achieved by introducing smaller planes on the Madrid–Barcelona route. As of March 2010, Iberia offered 30 flights a day on working days for each direction. The frequency offered by HSR was 27 trains a day for each direction, but only 10 of them were nonstop trains. This makes the HSR frequency less appealing for users than the PA frequency.

### *Reliability*

One main advantage of HSR services compared with air transportation is punctuality and reliability. Actually, the percentage of PA flights getting to their destination on time in 2010 was 92% compared with 99.3% for HSR. HSR services are less sensitive to weather conditions and congestion problems than is air transportation. Moreover, in the case of a delay longer than 15 min RENFE reimburses 50% of the ticket price; the reimbursement is 100% if the delay is longer than 30 min. PA does not offer any kind of reimbursement for delay.

### *Comfort and Possibility of Additional Activities During Trip*

Another crucial advantage of HSR compared with PA is greater comfort for the user and the possibility of taking advantage of travel time

to work or to engage in other activities. HSR is more comfortable in regard to space for users. The seats are wider and the distance between seat rows is longer (90 cm in the economy class of HS trains versus 73 cm in the economy class of the air shuttle). Inside the trains, users may have access to the Internet, and they can use their cell phones. Moreover, the train is equipped with a cafeteria car where passengers can have meals or drinks. Because of those aspects, users perceive HSR travel to be better than traveling by plane.

## **MADRID–BARCELONA CORRIDOR AFTER HSR**

### **Survey**

The results of a survey conducted between February and March of 2010 have been used to analyze the competition between HSR and PA along the Madrid–Barcelona corridor. The reference universe is made up of all users who in the reference period traveled along the corridor with HSR and PA, moving from the Atocha train station and Barajas Airport in Madrid, respectively. The survey was designed specifically to analyze the demand response to additional policies—different from the obvious and substantial reduction in travel time after the opening of the HSR service—which would affect the competitiveness of both modes in the near future. Therefore, specific questions were included to obtain information about other important service attributes.

The questionnaire submitted to users was made up of two parts: the first included RP questions concerning users' socioeconomic characteristics, trip purpose, time at destination, travel class, and so on (Table 2). The second part collected the information needed to specify and then to calibrate a mode choice model between HSR and PA. Although the survey did not contain a standard stated choice experiment, some SP questions were included to analyze the effect of travel cost, service frequency, parking availability at the train station, as well as the ease of security controls at the airport.

From the RP survey, it has been found that most users are men, 71% and 75% from Atocha and Barajas, respectively. Most users are between 36 and 50 years old. The income level ranged between €40,000 and €80,000 a year. The main trip purpose was work for 66% of the users traveling from Barajas and 81% for those who traveled from Atocha. The specific details of the sample distribution are shown in Table 3. Concerning the SP survey, users were

**TABLE 2** Characteristics of Questionnaire

#### Revealed Preferences Variable

##### Socioeconomic characteristics

1. Gender
2. Age (years)
3. Residential place
4. Destination place
5. Income level (€/year)
6. Number of household components

##### Travel information

7. Trip purpose
8. Time at destination (both in Madrid and Barcelona)
9. Travel class
10. Travel time to reach terminal from residential place
11. Mode chosen to reach terminal
12. Mode to reach final destination from terminal
13. Which is the most expensive transport mode along this corridor?
14. Please specify the factors positively influencing the mode choice.

TABLE 3 Descriptive Analysis of Sample

Category	Barajas (%)	Atocha (%)
Age (years)		
18–24	10	6
25–35	39	26
36–50	35	44
51–65	14	18
>65	10	6
No response	1	2
Income (€/year)		
0–20,000	10	6
20,000–40,000	39	26
40,000–80,000	35	44
80,000–150,000	14	18
>150,000	10	6
No response	1	2
Trip purpose		
Work	66	81
Leisure	25	7
Visiting relatives or friends	4	9
Other purposes	5	3
Travel class		
Economy	98	81
Business	2	18
Club	0	1

presented with different scenarios representing possible changes in the transportation supply system in regard to the following attribute changes:

1. Increase in fare ticket for competing mode (in Barajas and Atocha),
2. Increase in service frequency of HSR (in Barajas),
3. Improved parking opportunities at the train station (in Barajas), and
4. Eased security controls at the airport (in Atocha).

### Mode Choice Model

Discrete choice models have been widely used to study travelers' behavior in the mode choice context. The theoretical underpinnings are found in the theory of rational choice and in the utility maximization behavioral rule. Thus, the utility to the decision maker is represented by the random variable  $U_{jq} = V_{jq} + \varepsilon_{jq}$ , where  $V_{jq}$  is the deterministic or observable utility and  $\varepsilon_{jq}$  is a random term representing the portion of utility unknown to the analyst. Therefore, under the assumption of utility maximization, it is possible to model only the choice probability of the different alternatives.

Different assumptions about the distribution of the unobserved portion of utility  $\varepsilon_{jq}$  result in different representations of the choice model. Thus, the widely used multinomial logit (MNL) and nested logit (NL) models are obtained when  $\varepsilon_{jq}$  are independent and identically distributed (iid) extreme values and a type of generalized extreme value, respectively [see Train (18) and Ortúzar and Willumsen (19) for more details about the derivation of the choice probabilities]. The mixed logit (ML) model solves the main limitations of the MNL and NL models. The ML model allows for random taste variation, unrestricted substitution patterns, and even correlation in unobserved factors over time, which is particularly useful when one deals with

SP or panel data. The ML model is a very flexible model that can approximate any random utility model with total precision (20). Under the random coefficient version, the utility of alternative  $j$  for an individual  $q$  is represented by  $U_{jq} = \beta'_q x_{jq} + \varepsilon_{jq}$ , where  $x_{jq}$  is a vector of observed attributes of alternative  $j$  for decision maker  $q$ ,  $\varepsilon_{jq}$  is a set of random variables iid extreme value, and  $\beta_q$  is a vector of random coefficients. In the error component formulation of the ML model, utility is represented by  $U_{jq} = \alpha' x_{jq} + \mu'_q z_{jq} + \varepsilon_{jq}$ , where  $x_{jq}$  and  $z_{jq}$  are vectors of observed attributes of the alternative  $j$  for individual  $q$ ,  $\alpha$  is a vector of fixed coefficients,  $\mu_q$  is a vector of random terms with zero mean and covariance, and  $\varepsilon_{jq}$  are defined as above.

Three different models, based on the following linear-in-the-parameter specification for utility, were considered for this data set:

$$\begin{aligned}
 V_{\text{AIR}} &= \beta_{\text{ASC\_AIR}} + \beta_{\text{COST}} \cdot \text{COST}_{\text{AIR}} + \beta_{\text{FREQ}} \cdot \text{FREQ}_{\text{AIR}} \\
 &\quad + \beta_{\text{CHECK-IN}} \cdot \text{CHECK-IN}_{\text{AIR}} \\
 V_{\text{AVE}} &= \beta_{\text{COST}} \cdot \text{COST}_{\text{AVE}} + \beta_{\text{FREQ}} \cdot \text{FREQ}_{\text{AVE}} \\
 &\quad + \beta_{\text{PARKING}} \cdot \text{PARKING}_{\text{AVE}} \cdot \text{CAR}
 \end{aligned} \quad (1)$$

where

COST = travel cost in euros,

FREQ = service frequency measured in departures per hour,

CHECK-IN = 1 if security control service and check-in at the airport are rapid and smooth and 0 otherwise,

PARKING = 1 if parking capacity at the train station is good and 0 otherwise,

CAR = 1 if access is by car and 0 otherwise, and

$\beta$ s = unknown parameters.

In particular,  $\beta_{\text{ASC\_AIR}}$  represents the air-alternative-specific constant, while other parameters represent the marginal utility of the corresponding attributes. Because particular interest lies in analyzing the policy consisting of improving parking facilities at the Atocha train station, the study of this effect refers only to trips with a home end in Madrid.

For the first model (MNL1), a binomial logit model specification was considered; for the second model (ML2) an error component panel mixed logit model was considered with fixed parameters but accounting for the correlation between the choices made by the same individual. Finally, in the third model (ML3), random taste heterogeneity was analyzed considering the specification of a random cost parameter following the normal distribution. Random taste heterogeneity was also tested for the rest of the parameters, but no consistent specification was found with more than two random parameters.

Unknown parameters were calibrated with BIOGEME 1.8 (21). Estimated results are shown in Table 4.

In all cases, parameters were significant at the 95% confidence level and presented the expected sign. The only exception was the standard deviation of the cost parameter, which was significant at the 94% confidence level. An increase in the cost of the ticket makes both alternatives less attractive as does a decrease in service frequency. A hypothetical increase in parking capacity around Atocha Station would make the HSR alternative more attractive, and the same would be true for the air alternative after a reduction in the time relative to check-in and security control at Barajas Airport. In that respect, the PARKING variable was significant only for those passengers who used a private car as their access mode to the

TABLE 4 Estimation Results: After HSR

Attribute	MNL1		ML2		ML3	
	Parameter Estimate	t-Test	Parameter Estimate	t-Test	Parameter Estimate	t-Test
ASC_AIR	7.3500	11.54	11.0000	8.56	11.8000	7.32
CHECK-IN	1.3300	4.41	2.2600	5.16	2.4000	4.57
COST	-0.0696	-11.85	-0.1030	-8.59	-0.1120	-7.37
COST_SIGMA	—	—	—	—	0.0135	1.86
FREQ	0.6610	11.01	0.9650	8.12	1.0900	6.29
PARKING*CAR	0.5950	2.01	1.2600	2.84	1.4500	2.64
SIGMA	—	—	1.3700	5.54	1.4500	4.98
$L^*(0)$	-700.772		-700.772		-700.772	
$L^*(C)$	-689.752		-689.752		-689.752	
$L^*(\Sigma)$	-466.925		-456.638		-453.662	
$\rho^2$	.334		.348		.353	
$\rho^2$ adjusted	.327		.340		.343	
Observations	1,011		1,011		1,011	

NOTE: — = not applicable.

airport; that is, when the interaction PARKING\*CAR was specified. Because the specification of the model does not include travel time as an explanatory variable, the effect of this attribute is confused with the alternative-specific constant for the air alternative, which explains the high value obtained for this parameter. Therefore, the positive sign of this constant should be interpreted in this case as passenger preference, in regard to total travel time and other unobserved factors, for the air alternative when the effect of the rest of the attributes is negligible.

The standard deviation (SIGMA) of the error component was found to be significant in both the ML2 and the ML3 models, thus verifying the existence of a correlation between choices made by the same individual in the SP questions. The size of this parameter indicates that the correlation is higher in model ML3. Because this model also presented a better likelihood and a better goodness of fit (according to the value of the rho square indices), it is preferred over the other two models.

## Demand Analysis

From the estimates carried out, some results can be obtained on what effects changes in the supply system would have on users' choices. Assuming a 3% increase in total traffic (to 5.815 million passengers) for 2011, the model predicts a 44.15% market share for HSR in this corridor given the actual situation. Improvements in the parking facilities at Atocha Station do not represent a substantial increment in the market share of HSR (46.11%). However, improvements in check-in and security control processes at Barajas Airport would place the plane in a dominant position (77.64%) compared with that enjoyed by this mode before the entrance of HSR in the corridor. Departing from the actual situation, for HSR to achieve 50% of the market, a fare reduction of 3.76% would be needed. If the objective for the rail operator is to obtain more than a 10% advantage over the competing mode, a fare reduction of 6.63% should be accompanied by an incremental increase in the service frequency of 1.29%. A higher effort would be required to exert

a strong dominance (more than 60% of the share) in this market. In this case fares should decrease by 9.55% and frequency should increase by 1.81% (Table 5). These results reinforce the importance of prices and service frequency as the main instruments of modal competition for HSR in this market. The effect of other policies, such as improvement in parking facilities at the train station, would play a definitively secondary role.

## Willingness to Pay for Improving Level of Service

Willingness to pay (WTP) measures, in monetary terms, express changes in the utility resulting from changes in the level-of-service attributes. In other words, the WTP is represented by the marginal rate of substitution between travel cost and the corresponding attribute. WTP measures are derived from estimates of discrete choice

TABLE 5 Demand Responses to Improvements in Supply System

Policy Scenario	Market Share		Passengers (thousands)	
	Plane (%)	HSR (%)	Plane	HSR
Actual situation	55.85	44.15	3,248	2,568
Improvement in parking	53.89	46.11	3,134	2,681
Improvement of check-in and security process in Barajas Airport	77.64	22.36	4,515	1,300
HSR fare reduction: 3.76%	50.00	50.00	2,908	2,908
HSR fare reduction: 6.63%; HSR frequency increment: 1.29%	45.00	55.00	2,617	3,198
HSR fare reduction: 9.55%; HSR frequency increment: 1.81%	40.00	60.00	2,326	3,489

models as the ratio between the marginal utility of this attribute ( $q_{kj}$ ) and the marginal utility of the travel cost ( $c_j$ ):

$$WTP_{q_{kj}}^j = -\frac{dc_j}{dq_{kj}} = \frac{\frac{\partial V_j}{\partial q_{kj}}}{\frac{\partial V_j}{\partial c_j}} \quad (2)$$

where  $V_j$  is deterministic or observable utility of the alternative  $j$  as introduced previously.

In the case of qualitative variables, the WTP for improving an attribute (e.g., passing from Level 0 to Level 1) is given by the following expression:

$$WTP_{q_{kj}}^j = \frac{V_j^1 - V_j^0}{\lambda} \quad (3)$$

where  $V_j^0$ ,  $V_j^1$  represent the observable utility when the attribute takes Level 0 and 1, respectively; and  $\lambda$  is the marginal utility of income, which coincides with minus the marginal utility of the travel cost ( $-\partial V_j / \partial c_j$ ).

When the specification of  $V_j$  is linear with fixed parameters, as in MNL1 and ML2 models, Equation 2 yields the quotient between the coefficients of  $q_{kj}$  and travel cost. Therefore, the point estimate of the WTP is represented by a fixed value. In the ML3 model, because the cost parameter is a random variable, the WTP measures for this model are random variables as well. Therefore, computation of the WTP distributions requires simulation of the cost parameter normal distribution according to the point estimates for the mean and standard deviation. The table below shows the WTP measures obtained for models considered in the analysis. Values reported for the ML3 model below correspond to the WTP computed at the mean value of the cost parameter. In general, mixed logit specifications exhibit higher values for the WTP measures. The only exception is found for the service frequency in the ML2 model.

Attribute	MNL1	ML2	ML3
Check-in (€)	19.11	21.94	21.43
Frequency (€/service)	9.50	9.37	9.73
Parking*Car (€)	8.55	12.23	12.95

The highest WTP, ranging from €19 to €22, is found for speeding up the check-in and security control processes at the airport. The three models presented a very similar WTP figure for having an additional departure per hour (about €9). Finally, the WTP for parking space availability at the train station ranges from €8 to €13, depending on the specification considered.

Figure 2 depicts the distribution of the WTP measures for the ML3 model after consideration of 10,000 random draws of the normal distribution for the cost parameter. The 95% confidence intervals for these distributions are also reported. Results show that the WTP for greater frequency presented the distribution with the lowest variance. However, as the only differences in the shape of the distribution are due to changes in the scale of the variable, all of the distributions presented the same value for the coefficient of variation: 12.61%.

## CONCLUSIONS

This paper shows the big impact that introducing an HSR service between two big cities with a strong business relationship, such as is the case for Madrid and Barcelona, may have on the modal share distribution in the corridor. Despite that, the ultimate results show that the market share taken on by HSR has been lower than what was originally predicted. The difference was probably caused by the fact that modelers did not take into account the response of the air industry, especially PA, to the opening of HSR services, which involved reducing prices and maintaining frequency with smaller planes.

PA and HSR are presently engaged in a strong fight to get the lion's share in the corridor. HSR is more comfortable for users because they do not have to go through security checkpoints and can make better use of their time inside the train. Moreover, HSR stations are on average more accessible for users than are airports, particularly for those users who take public transportation to get to or

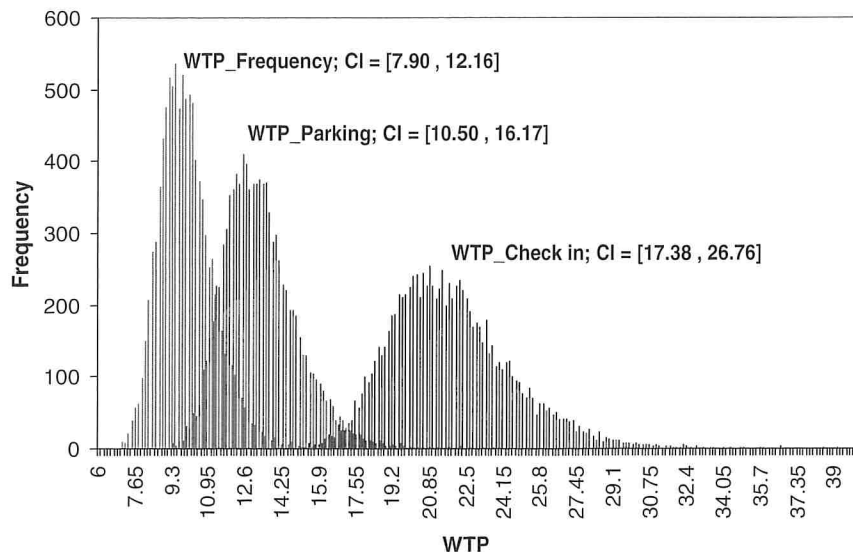


FIGURE 2 Distribution of WTP measures for ML3 model (CI = confidence interval).

to leave from the station or airport. In spite of this, PA still maintains important advantages over HSR. PA's frequency is unbeatable. HSR has fought to get a frequency similar to that of PA, but most of the trains stop along the way, making travel times longer. Moreover, PA is more attractive for users who need to use a car to get to the airport or station.

The major findings of this paper are focused on the attempt to know the demand response to different policy scenarios designed with the purpose of improving the level of the transport services. Estimated results based on the SP sample show that prices and service frequency are still the main instruments that both HSR and PA can use to compete against alternative modes in this corridor. The models confirm the positive effect of improving parking facilities at the train stations. A hypothetical increase in the parking capacity around Atocha Station would make the HSR alternative more attractive. However, the demand analysis carried out demonstrates that the effect of this measure will be low. By contrast, the ease of the check-in and security control processes at the airport, the variable having the highest WP, would represent a substantial increase for air transport demand. The hypothetical avoidance of check-in and security control processes at Barajas Airport would place the plane in a dominant position (77.64%) compared with that enjoyed by the plane mode before the entrance of HSR into the corridor. Notwithstanding, it is acknowledged that other effects, such as the airline's response in reducing fares and the size of the planes to maintain the level of service frequency, as well as the time penalty produced by extra stops in the HSR services, could also represent important instruments of competition in this market.

The main policy lesson from this research is that, apart from travel time, there are other crucial aspects to make the HSR mode competitive with air transportation. Frequency and price are likely the most important. However, the potential to reduce prices and increase frequency for HSR while preserving the financial sustainability of the project requires a high level of potential demand in the corridor. In other words, corridors that do not have enough demand to enable a certain service frequency will rarely be able to compete with air transportation.

## REFERENCES

1. Vickerman, R. High-Speed Rail in Europe: Experience and Issues for Future Development. *Annals of Regional Science*, Vol. 31, 1997, pp. 21–38.
2. Bruinsma, F., and P. Rietveld. Urban Agglomerations in European Infrastructure Networks. *Urban Studies*, Vol. 30, 1993, pp. 919–934.
3. Blum, U. Markets for High-Speed and Fast Trains: Development Patterns in Germany and Europe. Presented at Workshop on Regional and Urban

- Effects of High-Speed Trains, Jönköping International Business School, Jönköping, Sweden, 1995.
4. *Plan Estratégico de Infraestructuras y Transportes*. Ministerio de Fomento, Madrid, Spain, 2005.
5. de Rus, G., and C. Román. Análisis económico de la línea de alta velocidad Madrid-Barcelona. *Revista de Economía Aplicada*, Vol. 14, No. 42, 2006, pp. 35–79.
6. de Rus, G., and G. Nombela. Is the Investment in High Speed Rail Socially Profitable? *Journal of Transport Economics and Policy*, Vol. 41, No. 1, 2007, pp. 3–23.
7. Martín, J. C., and G. Nombela. Microeconomic Impacts of Investments in High Speed Trains in Spain. *Annals of Regional Science*, Vol. 41, No. 3, 2007, pp. 715–733.
8. Gutiérrez, J., R. González, and G. Gómez. The European High-Speed Train Network: Predicted Effects on Accessibility Patterns. *Journal of Transport Geography*, Vol. 4, No. 4, 1996, pp. 227–238.
9. Gutiérrez, J. Location, Economic Potential and Daily Accessibility: An Analysis of the Accessibility Impact of the High-Speed Line. *Journal of Transport Geography*, Vol. 9, 2001, pp. 229–242.
10. Martín, J. C., J. Gutiérrez, and C. Román. Data Envelopment Analysis (DEA) Index to Measure the Accessibility Impacts of New Infrastructure Investments: The Case of the High-Speed Train Corridor Madrid-Barcelona–French Border. *Regional Studies*, Vol. 38, No. 6, 2004, pp. 697–712.
11. González-Savignat, M. Competition in Air Transport. The Case of the High Speed Train. *Transport Reviews*, Vol. 24, No. 3, 2004, pp. 293–316.
12. González-Savignat, M. Will the High-Speed Train Compete Against the Private Vehicle? *Journal of Transport Economics and Policy*, Vol. 38, No. 1, 2004, pp. 77–108.
13. López-Pita, A., and F. Robusté. Impact of High-Speed Lines in Relation to Very High Frequency Air Services. *Journal of Public Transportation*, Vol. 8, No. 2, 2005, pp. 17–35.
14. Román, C., R. Espino, and J. C. Martín. Competition of High Speed Train with Air Transport: The Case of Madrid-Barcelona. *Journal of Air Transport Management*, Vol. 13, No. 5, 2007, pp. 277–284.
15. Román, C., R. Espino, and J. C. Martín. Analyzing Competition Between the High Speed Train and Alternative Modes. The Case of the Madrid-Zaragoza-Barcelona Corridor. *Journal of Choice Modeling*, Vol. 3, No. 1, 2010, pp. 84–108.
16. Román, C., and J. C. Martín. Potential Demand for New High Speed Rail Services in High Dense Air Transport Corridors. *International Journal of Sustainable Development and Planning*, Vol. 5, No. 2, 2010, pp. 114–129.
17. Román, C., and J. C. Martín. The Effect of Access Time on Modal Competition for Interurban Trips: The Case of the Madrid-Barcelona Corridor in Spain. *Networks and Spatial Economics*, Vol. 11, No. 14, 2011, pp. 661–675.
18. Train, K. *Discrete Choice Methods with Simulation*. Cambridge University Press, Cambridge, United Kingdom, 2003.
19. Ortúzar, J. D., and L. G. Willumsen. *Modelling Transport*, 3rd edition, John Wiley & Sons, Chichester, United Kingdom, 2001.
20. McFadden, D., and K. Train. Mixed MNL Models of Discrete Choice Response. *Journal of Applied Econometrics*, Vol. 15, 2000, pp. 447–470.
21. Bierlaire, M. *Estimation of Discrete Choice Models with BIOGEME 1.8*. École Polytechnique Fédérale de Lausanne EPFL, 2009.

---

*The Intercity Passenger Rail Committee peer-reviewed this paper.*



# High-Speed Route Improvement Optimizer

Yung-Cheng (Rex) Lai and Po-Wen Huang

The most common near-term approach to upgrading railway corridors to achieve high speeds is by incremental improvement. Existing infrastructure or rolling stock can be improved in various ways to allow for increasing speeds and reducing travel time along a route. Each track section has a characteristic set of opportunities for increasing speed and their corresponding effects on travel time reduction, along with an associated set of costs. Similarly, each type of rolling stock has the potential to increase operational speed and reduce travel time, corresponding to a specific price tag. This research focuses on developing a decision support process by using mathematical programming to identify the most cost-effective strategy for reducing corridor travel time given a prescribed performance goal and budget. This optimization process was implemented in a railway corridor in Taiwan, and the result showed that the best strategy was a combination of infrastructure upgrade and rolling stock acquisition. Using this tool can help states and railway agencies simultaneously maximize capital investment returns and maintain reliable services for their customers.

To provide faster and more reliable intercity services, high-speed rail (HSR) is a clear choice for future intercity transportation (1–3). HSR can generally be achieved through two different approaches: (a) an incremental approach, which involves improving existing infrastructure, and (b) dedicated lines, which involve constructing new dedicated high-speed infrastructure (4, 5). Constructing a dedicated high-speed line can significantly reduce travel time. However, a dedicated line is costly, involves political difficulties, and requires a lengthy time frame for project completion. Improvements from the incremental approach may be modest but they are completed relatively rapidly, and the initial cost is much lower compared with the cost of the alternatives. As a result, in several countries, such as the United States and Sweden, the most common near-term approach to upgrading railway corridors to achieve high speeds is incremental improvement (6, 7).

The railway system in Taiwan presents a similar scenario. After the completion of Taiwan High Speed Rail in 2007, the construction of another high-speed line is highly unlikely. To provide more efficient and reliable transportation between the capital, Taipei, and eastern Taiwan, upgrading the conventional railway has become the focus of the Railway Reconstruction Bureau (RRB). Along with the operator of the conventional railway, the Taiwan Railways Administration (TRA) and RRB are currently looking into the possibility of reducing the travel time between Nangang and Toucheng (part

of the Yilan Line) (see Figure 1) to improve their market share in transportation from Taipei to eastern Taiwan (8, 9). TRA is facing a highly competitive service provided by the intercity bus sector, and a 16-min reduction in travel time can boost the demand by about 10% (9, 10). Figure 1 shows the existing and planned lines for this corridor. The planned line consists of a set of alternatives to upgrade the infrastructure via an incremental approach (8).

Through the incremental approach, existing infrastructure or rolling stock can be improved in various ways to increase train speed, or reduce travel time, or both. Such enhancements may include track quality upgrades, improved alignments, signal and traffic control system modifications, and special train set acquisition. Each track section has a characteristic set of opportunities for increasing speed and their corresponding effects on travel time reduction. Similarly, each type of rolling stock has the potential to increase operational speed to reduce travel time (Table 1) (11–14).

The conventional route improvement process still requires human intervention and expert judgment. For decades, simulation tools have been widely used to assist in the strategic capacity planning process (15–17). These tools can simulate a corridor of the network in a more detailed way (18, 19), but because of computational constraints they are not suitable for network analysis with a list of possible alternatives for every link (20, 21). To tackle this complexity, several studies used network optimization techniques to develop decision support tools for the railway capacity planning process (21–25). However, none of these studies takes into account the possibility of improving services by investing in rolling stock, nor is the journey time considered in the optimization process.

HSR route studies often focus on increasing maximum speed on a route, but in many circumstances this may not result in the most cost-effective improvement in travel time (7, 26). There are practical constraints on speed at many locations, such as curves, grades, and bridges, and there are other factors that have a greater effect on travel time. Improving these factors may yield more benefit than modifications that allow very high speed operation in other areas because the marginal benefit in time saved is not as great as that of more modest improvements in speed-constrained areas. For example, increasing speeds from 70 km/h to 100 km/h on a segment will result in a 30% reduction in travel time, whereas upgrading an identical-length segment from 130 km/h to 160 km/h will reduce travel time by only 18%. The latter upgrade may also be more costly than the former, despite yielding a lower reduction in travel time. Therefore, establishing a decision support process is essential to identifying the most cost-effective strategy for reducing corridor travel time given a prescribed performance goal and budget.

In this study, a route improvement optimizer (RIO) is developed to identify the most cost-effective investment selections with the consideration of feasible improvement alternatives as well as quantifiable costs and benefits. On the basis of information on the current network and traffic and the available investment options, the RIO is able to successfully determine the optimal solution

Department of Civil Engineering, National Taiwan University, Room 313, Civil Engineering Building, No. 1, Roosevelt Road, Section 4, Taipei 10617, Taiwan. Corresponding author: Y.-C. Lai, yclai@ntu.edu.tw.

*Transportation Research Record: Journal of the Transportation Research Board*, No. 2289, Transportation Research Board of the National Academies, Washington, D.C., 2012, pp. 18–23.  
DOI: 10.3141/2289-03

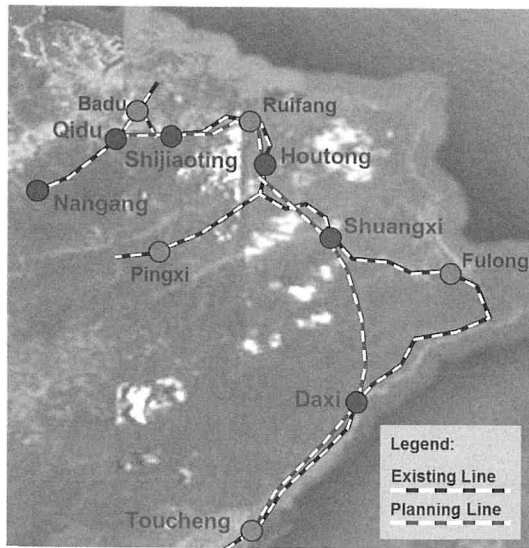


FIGURE 1 Proposal for Yilan Line upgrade alternatives (8).

to which sections should be upgraded, the type of engineering options that should be implemented, and the best rolling stock option. The RIO is also implemented to identify the optimal set of route improvements for the Yilan Line. Through the RIO, states and railway agencies will be able to simultaneously maximize capital investment returns and maintain reliable services for their customers.

## DECISION SUPPORT PROCESS FOR ROUTE IMPROVEMENT

Figure 2 shows the proposed decision support process for route improvement. According to the infrastructure properties and alternatives provided by users, the network segmentation process first defines the appropriate nodes and links of the corridor for the RIO module. This information, together with the rolling stock properties and alternatives, is then used to determine the infrastructure cost of upgrading the infrastructure, the running time on each link for each type of rolling stock under possible route configurations, and the cost of the rolling stock. The RIO optimization module then takes into account all possible route improvement options along with their costs and benefits to determine the optimal investment plan for selecting infrastructure projects and rolling stock.

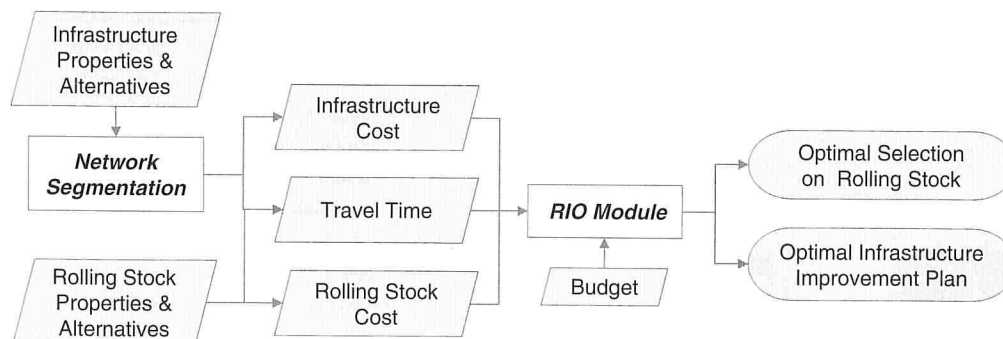


FIGURE 2 Route improvement optimization process.

TABLE 1 Speed Limit of Each Type of Train in Various Curvatures (13, 14)

Radius of Curvature (m)	Speed Limit in Curvature (km/h)		
	General Multiple Unit Train	Roller-Type Tilting Train	Air-Spring Tilting Train
0	130	130	130
700	110	130	130
600	100	125	120
500	90	115	105
450	85	110	100
400	80	100	95
350	75	95	85
300	70	90	80
250	65	85	75

## Network Segmentation and Alternatives Generation

The network segmentation process defines the appropriate nodes and links of the corridor. Using the existing stations and sections as the nodes and links in the optimization process is an intuitive approach. However, some sections may have multiple alternatives, and some alternatives may influence several sections. For example, Figure 3 shows a section between Houtong and Shuangxi. Three alternatives are available for this section: upgrading the link between any or all of the following: (a) Houtong and Sandiaoling, (b) Sandiaoling and Mudan, and (c) Mudan and Shuangxi. For this section, combining the upgrade of any of the three sections or treating them separately may be the optimal solution. This section is therefore separated into three subsections in the optimization process instead of using only one section (from Houtong to Shuangxi). Figure 4 shows the end result of the segmentation process for the Yilan Line. Compared with the original network structure in Figure 1, the revised network has 11 sections, and each is associated with a decision: to upgrade or not upgrade. The cost of upgrading each section is listed in Table 2.

Upgrading infrastructure is a common approach, but adopting trains with faster train sets is also a popular option in reducing travel time (26). For example, adopting tilting trains is a possible and popular option for a network with a considerable number of curves (Table 1). Improvements in infrastructure provide location-specific benefits, whereas improvement in the rolling stock offers location-free benefits. Table 3 shows the cost of possible improvements in the



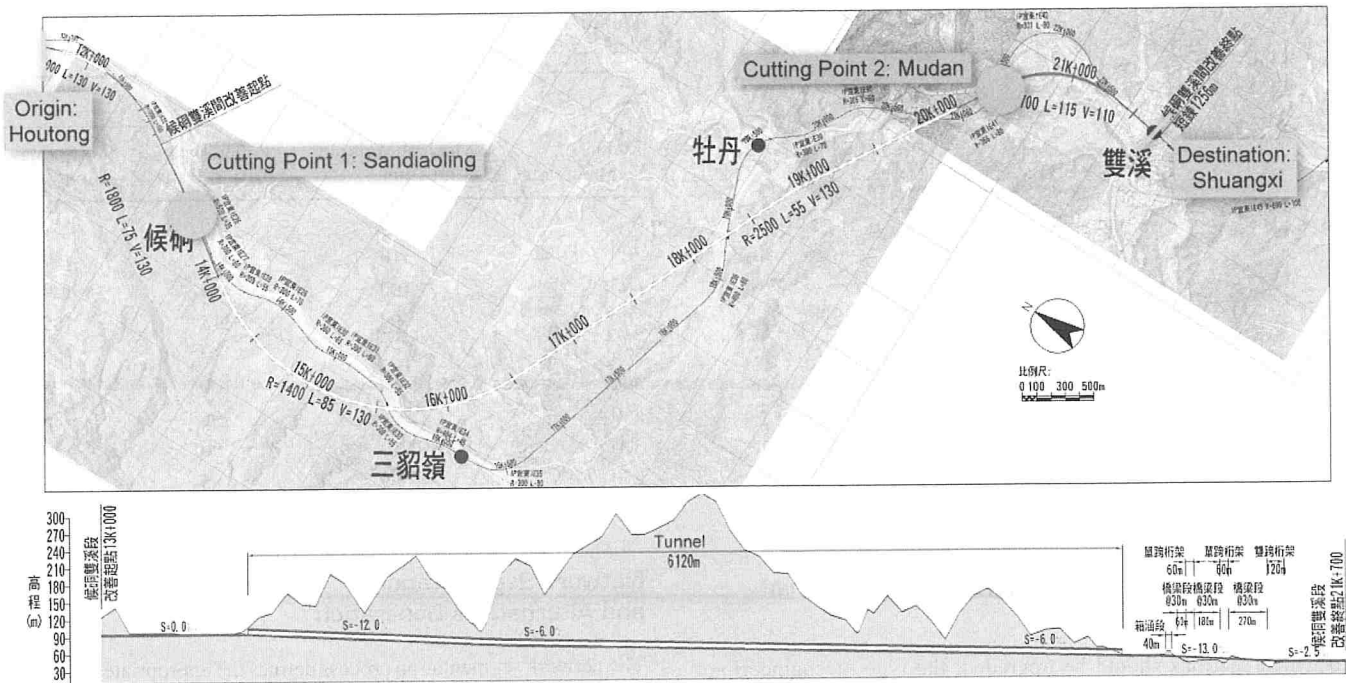


FIGURE 3 Upgrade proposal for Houtong to Shuangxi segment on Yilan Line (8).

rolling stock for the Yilan Line. New passenger services are assumed to require 48 cars, and the RIO should choose the most cost-effective train to offer these services (27). The push-pull type is not equipped with tilting technology and is the most inexpensive option. The roller-type tilting trains and air-spring tilting trains are more expensive, with the possibility of increasing speed through curves. The roller-type tilting train, the fastest rolling stock through curves, is the most expensive of the three types (Table 3). Therefore, there is a trade-off between speed and cost. Considering an infrastructure and a rolling stock upgrade, there is another trade-off between location-based improve-

ment and location-free improvement. The RIO module aims to take these factors into account and determine the optimal investment plan.

### RIO Optimization Module

The objective of the RIO is to minimize travel time on a limited budget. The objective can be formulated as a knapsack problem in resource allocation with financial constraints. The original knapsack problem determines the optimal collection of items, each with specific weight and value, that can be placed in a fixed-size knapsack. In this application, the possible infrastructure improvement and rolling stock alternatives represent the available items, and the limited budget represents the fixed-size knapsack. Instead of a collection of items, the RIO optimization module determines the most cost-effective investment strategy. This budget refers specifically to budget for capital investment regardless of the maintenance budget.

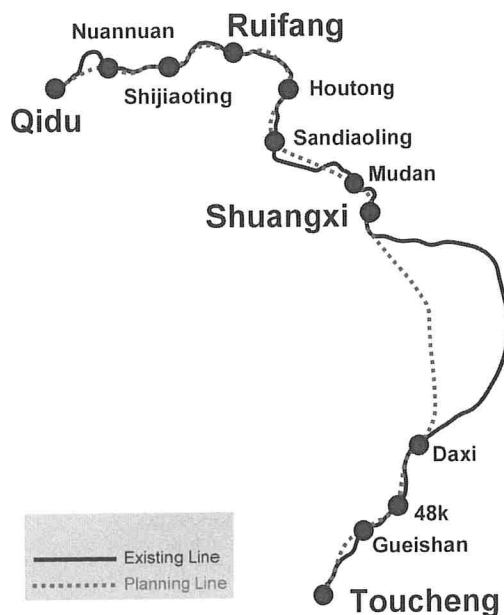


FIGURE 4 Study network (Yilan Line) (8).

TABLE 2 Sections with Corresponding Upgrade Costs

Section	Origin	Destination	Cost (NTD billions)
1	Qidu	→ Nuannuan	1.35
2	Nuannuan	→ Shijiaoting	1.18
3	Shijiaoting	→ Ruifang	1.70
4	Ruifang	→ Houtong	1.61
5	Houtong	→ Sandiaoling	0.57
6	Sandiaoling	→ Mudan	3.70
7	Mudan	→ Shuangxi	0.73
8	Shuangxi	→ Daxi	7.06
9	Daxi	→ 48k	1.50
10	48k	→ Gueishan	1.00
11	Gueishan	→ Toucheng	1.97

TABLE 3 Cost of Possible Rolling Stock for Yilan Line

Characteristic	Push-Pull Train	Roller-Type Tilting Train	Air-Spring Tilting Train
Cost per single car (NTD billions)	0.05	0.08	0.06
Quantity	48	48	48
Total cost (NTD billions)	2.40	3.84	2.88

According to TRA, the maintenance cost for rolling stock and infrastructure is similar among different alternatives so can be ignored in this stage of decision making (27).

The following notations are used in the mixed integer programming model:

- $I$  = set of route sections, indexed by  $i$ ;
- $Q$  = set of alternatives for upgrading the infrastructure, indexed by  $q$ ;
- $K$  = set of possible train types, indexed by  $k$ ;
- $B$  = available budget for capital investment;
- $t_i^{kq}$  = reduction in travel time from  $k$ th train and  $q$ th infrastructure upgrade;
- $h_i^q$  = cost of upgrading section  $i$  via option  $q$ ; and
- $c_k$  = cost of  $k$ th train.

Two sets of decision variables are used in this model: (a)  $x_k$  is a binary variable for the decision on train type  $k$  and (b)  $y_i^{kq}$  is a binary variable that determines whether the alternative  $q$  and  $k$ th train are selected for section  $i$ . This optimization module can be formulated as follows:

$$\max \sum_{k \in K} \sum_{i \in I} \sum_{q \in Q} t_i^{kq} y_i^{kq} \quad (1)$$

subject to

$$\sum_{k \in K} \sum_{i \in I} \sum_{q \in Q} h_i^q y_i^{kq} + \sum_{k \in K} c_k x_k \leq B \quad (2)$$

$$\sum_{k \in K} x_k = 1 \quad (3)$$

$$\sum_{k \in K} \sum_{q \in Q} y_i^{kq} = 1 \quad \forall i \in I \quad (4)$$

$$\sum_{q \in Q} y_i^{kq} \leq x_k \quad \forall i \in I, k \in K \quad (5)$$

and

$$\begin{aligned} x_k &\in \{0, 1\} & \forall k \in K \\ y_i^{kq} &\in \{0, 1\} & \forall k \in K, i \in I, q \in Q \end{aligned} \quad (6)$$

The objective function in Equation 1 maximizes the sum of the reduction time in each section. Equation 2 is the budget constraint, that is, the capital investment for infrastructure upgrade and acquisition cost of the rolling stock must be lower than or equal to the budget. Equation 3 ensures that only one train type is selected. Similarly, Equation 4 ensures that the model selects at most one upgrade alternative in section  $i$ . Finally, Equation 5 connects both

types of decision variables on rolling stock selection and infrastructure upgrade.

Equation 1 aims toward a maximization of travel time reduction. If there is a specific goal for the reduction time (e.g., a 10-min reduction), denoted by  $T$ , the optimization model should be reformulated as

$$\min \sum_{k \in K} \sum_{i \in I} \sum_{q \in Q} h_i^q y_i^{kq} + \sum_{k \in K} c_k x_k \quad (7)$$

such that Constraints 3, 4, 5, 6, and

$$\sum_{k \in K} \sum_{i \in I} \sum_{q \in Q} t_i^{kq} y_i^{kq} \geq T \quad (8)$$

The objective function in Equation 7 now aims to minimize the capital investment while subject to a predefined level of travel time reduction (Equation 8). This formulation makes it possible to determine an optimal budget to achieve the reduction time goal.

## CASE STUDY

As previously mentioned, the RIO process on upgrading part of the Yilan Line located northeast of Taiwan was applied (8, 9). In the segmentation process, this corridor was divided into 11 sections with a set of possible improvement alternatives. Table 4, derived from Table 5, shows the reduction time of the possible alternatives against that of the conventional push-pull trains. Along with possible rolling stock options, there were more than six thousand possible investment strategies, and the most cost-effective one would be difficult to determine manually. Therefore, the proposed optimization process was applied to this problem, and it was coded in the General Algebraic Modeling System (GAMS) and solved by using CPLEX (optimization software package) (28).

This optimal process can determine the best alternative combination on the basis of a particular budget level. In this case study, quite a few budget levels were tested to evaluate their effects on the optimal solution (Table 6).

For scenarios with budget levels less than 5 billion new Taiwan dollars (NTD) (1 NTD = US\$0.03, in 2012), route improvement via better rolling stock is more cost-effective than upgrading the infrastructure. In this budget range, the higher the budget, the faster the rolling stock selected for travel time reduction. Only when the budget level is more than 5 billion NTD will the selection of options associated with the infrastructure upgrades selected be seen. For scenarios with budget levels at 5 billion and 13.5 billion NTD, the roller-type tilting train was selected with a few infrastructure upgrades. However, if the budget level is more than 19 billion NTD, the combination of an air-spring tilting train with a set of infrastructure upgrades is a better solution compared with the more expensive roller-type tilting train. These results can help planners make the final decision and maximize capital investment returns.

## DISCUSSION OF FINDINGS

Improving existing railroad infrastructure and rolling stock for HSR service requires a substantial investment. Each modification has associated benefits in reduced travel time, as well as specific costs. In general, the higher the maximum speed, the higher the

TABLE 4 Travel Time in Minutes According to Types of Rolling Stock and Sections With or Without Infrastructure Upgrades

Yilan Line Segment	Without Upgrade (min)			With Upgrade (min)		
	Push-Pull Train	Roller-Type Tilting Train	Air-Spring Tilting Train	Push-Pull Train	Roller-Type Tilting Train	Air-Spring Tilting Train
Qidu–Nuannuan	3.16	2.47	2.73	1.38	1.38	1.38
Nuannuan–Shijiaoting	1.79	1.41	1.53	1.15	1.15	1.15
Shijiaoting–Ruifang	2.60	2.07	2.21	1.62	1.62	1.62
Ruifang–Houtong	2.19	1.72	1.88	1.62	1.62	1.62
Houtong–Sandiaoling	0.90	0.72	0.75	0.69	0.69	0.69
Sandiaoling–Mudan	6.04	4.78	5.30	3.00	3.00	3.00
Mudan–Shuangxi	1.89	1.53	1.69	0.82	0.69	0.69
Shuangxi–Daxi	12.00	9.60	10.00	6.00	6.00	6.00
Daxi–48k	3.00	2.40	2.50	1.38	1.38	1.38
48k–Gueishan	1.13	0.95	0.98	0.92	0.92	0.92
Gueishan–Toucheng	2.00	1.66	1.71	1.62	1.62	1.62
Total running time	36.70	29.31	31.26	20.20	20.08	20.08

cost. However, at many locations there are practical constraints on speed such as curves, grades, bridges, or other factors that have a greater effect on travel time. Improving these factors may yield more benefit than modifications that allow very high speed operation in other areas, because the marginal benefit in time saved is not as great as that of more modest improvements in speed-constrained areas. Although this approach may result in lower maximum speed on a line, it would offer equal or shorter total travel time, and at a lower cost.

The developed decision support process can determine the most cost-effective strategy, including infrastructure and rolling stock upgrades, for reducing corridor travel time given a prescribed performance goal and budget. The process and optimization module can also be adapted for scenarios in which only one of the two types of alternatives, upgrading infrastructure or rolling stock, is considered. For example, if only an infrastructure upgrade is possible, the decision variable on train type would be assigned a fixed value (i.e.,  $x_1 = 1$ )

and the rest of the formulation would remain the same. The outcome of this research can support the development of new HSR corridors and enable better prioritization of resources and cost savings in design, construction, and operations.

## CONCLUSION

The most common near-term approach to upgrading railway corridors to achieve high speeds is through incremental improvement. This research has developed a decision support process for identifying the most cost-effective strategy for reducing corridor travel time given a prescribed performance goal and budget. This optimization process was implemented in a railway corridor in Taiwan. The result shows that the best strategy is a combination of infrastructure upgrade and rolling stock acquisition. This finding also demonstrates that HSR route studies should focus on savings in travel time instead of increas-

TABLE 5 Reduction Time in Minutes According to Types of Rolling Stock and Sections With or Without Infrastructure Upgrades

Yilan Line Segment	Without Upgrade (min)			With Upgrade (min)		
	Push-Pull Train	Roller-Type Tilting Train	Air-Spring Tilting Train	Push-Pull Train	Roller-Type Tilting Train	Air-Spring Tilting Train
Qidu–Nuannuan	0.00	0.69	0.43	1.78	1.78	1.78
Nuannuan–Shijiaoting	0.00	0.38	0.26	0.64	0.64	0.64
Shijiaoting–Ruifang	0.00	0.53	0.39	0.98	0.98	0.98
Ruifang–Houtong	0.00	0.47	0.31	0.57	0.57	0.57
Houtong–Sandiaoling	0.00	0.18	0.15	0.21	0.21	0.21
Sandiaoling–Mudan	0.00	1.26	0.74	3.04	3.04	3.04
Mudan–Shuangxi	0.00	0.36	0.20	1.07	1.20	1.20
Shuangxi–Daxi	0.00	2.40	2.00	6.00	6.00	6.00
Daxi–48k	0.00	0.60	0.50	1.62	1.62	1.62
48k–Gueishan	0.00	0.18	0.15	0.21	0.21	0.21
Gueishan–Toucheng	0.00	0.34	0.29	0.39	0.39	0.39
Total running time	0.00	7.38	5.44	16.49	16.64	16.64

TABLE 6 Optimal Results Under Various Available Budgets

Yilan Line Segment	Available Budget (NTD billions)						
	2.5 <sup>a</sup>	3 <sup>b</sup>	4 <sup>c</sup>	5 <sup>c</sup>	13.5 <sup>c</sup>	19 <sup>b</sup>	26 <sup>b</sup>
Qidu–Nuannuan	e1	e1	e1	e1	e2	e2	e2
Nuannuan–Shijiaoting	e1	e1	e1	e1	e1	e1	e2
Shijiaoting–Ruifang	e1	e1	e1	e1	e1	e2	e2
Ruifang–Houtong	e1	e1	e1	e1	e1	e1	e2
Houtong–Sandiaoling	e1	e1	e1	e1	e1	e1	e2
Sandiaoling–Mudan	e1	e1	e1	e1	e1	e2	e2
Mudan–Shuangxi	e1	e1	e1	e2	e2	e2	e2
Shuangxi–Daxi	e1	e1	e1	e1	e2	e2	e2
Daxi–48k	e1	e1	e1	e1	e1	e2	e2
48k–Gueishan	e1	e1	e1	e1	e1	e1	e2
Gueishan–Toucheng	e1	e1	e1	e1	e1	e1	e2
Total reduction time (min)	0	5.42	7.39	8.23	12.92	15.78	16.64

NOTE: e1 = without upgrade; e2 = with upgrade.

<sup>a</sup>Push-pull train.

<sup>b</sup>Air-spring tilting train.

<sup>c</sup>Roller-type tilting train.

ing maximum speed on a route. Using the RIO can help states and railway agencies maximize capital investment returns and maintain reliable services for their customers.

## ACKNOWLEDGMENTS

The authors are grateful to Christopher P. L. Barkan from the Rail Transportation and Engineering Center at the University of Illinois at Urbana–Champaign for his assistance on this research. This project was funded by the National Science Council of Taiwan.

## REFERENCES

1. *High-Speed Ground Transportation for America*. FRA, Sept. 1997.
2. *Railway Technology Development 15 Years Planning and 2015 Long-Term Planning Outline*. Ministry of Railways, Beijing, China, 2002.
3. Kitagawa, T. Extending the Shinkansen Network. *Japan Railway and Transport Review*, No. 40, March 2005, pp. 14–17.
4. De Cerreno, A. L. C., D. M. Evans, and H. Permut. *High-Speed Rail Projects in the United States: Identifying the Elements for Success*. Mineta Transportation Institute, San Jose, Calif., Oct. 2005.
5. Barkan, C. P. L. Technical Challenges to Development of High-Speed Passenger Rail in the United States of America. *Proc., 1st International Conference on Railway Engineering*, Beijing, 2010.
6. Andersson, E., H. V. Bahr, and N. G. Nilstam. Allowing Higher Speeds on Existing Tracks—Design Considerations of the X2000 Train for Swedish State Railways. *Journal of Rail and Rapid Transit*, Vol. 209, 1995, pp. 93–104.
7. *High Speed Passenger Rail, Future Development Will Depend on Addressing Financial and Other Challenges and Establishing a Clear Federal Role*. U.S. Government Accountability Office, Washington, D.C., March 2009.
8. *The Advance Planning of Yilan Line and the North-Link Line Improving Program of TRA*. Railway Reconstruction Bureau, Ministry of Transportation and Communications, Taipei, Taiwan, 2009.
9. *The Speed Improvement Planning Between Nangang and Hualien Section of TRA*. Railway Reconstruction Bureau, Ministry of Transportation and Communications, Taipei, Taiwan, 2010.
10. *Before and After Opening Analysis Program Summary Report of Freeway No. 5 Nagang Su-ao Section*. Taiwan Area National Expressway Engineering Bureau, Ministry of Transportation and Communications, Taipei, Taiwan, 2010.
11. Kawabe, K. Technologies of High Speed Train Running in Curves. In *Understanding Rolling-Stocks*. Gakken Holdings Co., Ltd., Tokyo, Japan, 2007, pp. 82–83.
12. Chiou, J. T. The Developing Trend of Japanese Tilting Train. *TRA Journal*, No. 301, 1997, pp. 10–21.
13. *The Visual Encyclopedia of Railroad*. Shinsei Publishing Co., Ltd., Taito, Japan, 2007.
14. *An Introduction to Mechanics*. Taiwan Railway Administration, Ministry of Transportation and Communications, Taipei, Taiwan, 1985.
15. *I-5 Rail Capacity Study*. HDR, Inc., Portland, Ore., 2003.
16. Kittelson and Associates, Inc. *TCRP Report 100: Transit Capacity and Quality of Service Manual*, 2nd ed. Transportation Research Board of the National Academies, Washington, D.C., 2003.
17. Vantuono, W. C. Capacity Is Where You Find It: How BNSF Balances Infrastructure and Operations. *Railway Age*, Feb. 2005.
18. Kikuchi, S. A Simulation Model of Train Travel on a Rail Transit Line. *Journal of Advanced Transportation*, Vol. 25, No. 2, 1991, pp. 211–224.
19. Bandara, J. M. S. J., and I. A. B. Ekanayake. Train Scheduling Simulation That Minimises Operational Conflicts due to Service Constraints. *Journal of Advanced Transportation*, Vol. 37, No. 2, 2003, pp. 211–230.
20. Krueger, H. Parametric Modeling in Rail Capacity Planning. *Proc., Winter Simulation Conference*, Phoenix, Ariz., 1999.
21. Lai, Y.-C., and C. P. L. Barkan. A Comprehensive Decision Support Framework for Strategic Railway Capacity Planning. *ASCE Journal of Transportation Engineering*, Vol. 137, No. 10, 2011, pp. 738–749.
22. Loureiro, C. F. G., and B. Ralston. Investment Selection Model for Multicommodity Multimodal Transportation Networks. In *Transportation Research Record 1522*, TRB, National Research Council, Washington, D.C., 1996, pp. 38–46.
23. Petersen, E. R., and A. J. Taylor. An Investment Planning Model for a New North-Central Railway in Brazil. *Transportation Research Part A*, Vol. 35, No. 9, 2001, pp. 847–862.
24. Delorme, X., J. Rodriguez, and X. Gandibleux. Heuristics for Railway Infrastructure Saturation. *Theoretical Computer Science*, Vol. 50, No. 1, 2001, pp. 39–53.
25. Lai, Y.-C., M.-C. Shih, and J.-C. Jong. Railway Capacity Model and Decision Support Process for Strategic Capacity Planning. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 2197, Transportation Research Board of the National Academies, Washington, D.C., 2010, pp. 19–28.
26. Schmid, F. Control and Operation of Tilting Train Services. *Journal of Rail and Rapid Transit*, Vol. 212, 1998, pp. 76–77.
27. *Statistical Yearbook of Taiwan Railway*. Taiwan Railway Administration, Ministry of Transportation and Communications, Taipei, Taiwan, 2005.
28. *GAMS—A User's Guide*. GAMS Development Corporation, Washington, D.C., 2008.

The Intercity Passenger Rail Committee peer-reviewed this paper.

# Decision Support System to Optimize Railway Stopping Patterns

## Application to Taiwan High-Speed Rail

Jyh-Cherng Jong, Chian-Shan (James) Suen, and S. K. (Jason) Chang

An intercity passenger rail is built to connect several major cities. To provide satisfactory services to passengers, railway operators plan different stop schedules, such as all-stop, skip-stop, and express services. However, stopping patterns determined by empirical rules or political arguments are generally not optimal. This paper aims at developing a decision support system to generate the optimal combination of stopping patterns for minimizing total passenger in-vehicle time. This problem was first formulated by using mixed integer programming, but this method is intractable when dealing with large-scale problems because of the complexity of model structure and the nature of the problem. A genetic algorithm was then developed to search for the optimal or near-optimal solution efficiently within a reasonable computation time. The proposed algorithm was successfully implemented on Taiwan High Speed Rail. The resulting solution is better than the current practice, and the proposed algorithm is capable of finding the optimal solution in seconds. The present case study demonstrates that the decision support tool can tackle large-scale problems and can help operators efficiently and effectively design an optimal combination of stopping patterns.

An intercity passenger rail is the traffic backbone that connects a country's major cities. The travel demands for different origin and destination pairs are heterogeneous over time and space. However, the resources and facilities of a railway are limited. Thus, preparing a good operation plan to meet demands by suitable allocation of resources is an important planning issue for railway operators (1).

Planning railway operations is a complex process. In common railway practices, this planning is usually divided into a sequence of decisions (2–4). Figure 1 presents a typical planning process for intercity railway systems. It shows that railway operators usually first forecast the origin–destination (O-D) matrix through marketing research, transportation demand modeling, or analysis of historical ticketing data. Train stopping patterns are then drafted to meet travel demands. Once the stopping patterns are specified, operators can determine the service frequency of each pattern in different time periods by considering capacity, cost, and other limitations. This determination is

called train service planning. Finally, train scheduling is carried out to resolve train conflicts. The resulting timetable can be further applied to other resource planning activities, such as rolling stock planning, crew scheduling and rostering, or track occupancy planning.

The planning process reveals that train stop planning plays an important role in linking the demand side and the supply side and should be determined at an early stage of railway operation planning. The commonly used stopping patterns include all-stop, skip-stop, and express services (5, 6). Because the number of possible stopping patterns is enormous but only a few are selected for operations, it is difficult to obtain an optimal combination of stopping patterns by simple calculation. In common railway practice, empirical rules or sometimes political pressure may determine train stopping plans. These compromised plans are shortsighted and may increase passenger travel time (7).

In regard to train service planning, most studies assumed that train stopping patterns are a predetermined input (8–15). In other words, only a small set of possible stopping patterns were considered in determining service frequencies. Thus, the optimality cannot be guaranteed. Quite a few studies combined train stop planning with train service planning. For example, Chang et al. demonstrated a multi-objective model to optimize train frequencies while deciding stopping patterns dynamically (16). Ulusoy et al. formulated a nonlinear mixed integer programming (MIP) model to simultaneously obtain the optimal stopping patterns and service frequencies (17). Because this kind of integrated optimization model was very complicated and difficult to solve, the planning horizon was restricted to only 1 h. Thus, solving a full-range and more realistic problem is still a challenge to railway operators. Moreover, the most significant deficiency in existing train service planning models is that almost all models assume that trains departing from their origins can serve demands at any downstream station during the same hour. This assumption implies that train speed goes to infinite and the demands are uniformly distributed within 1 h. This outcome is generally not true because the travel time of an intercity train is usually more than 1 h, and the demands at far downstream stations are impossible to be served by the trains dispatching from their origins in the same hour. The only exception would be the study done by Jong and Suen, in which static passenger demands were adjusted according to train service speeds (10). However, in Jong and Suen's study, train stopping patterns were a predetermined input to the optimization model without justification (10). Thus, a good train stop planning model that can determine an optimal combination of stopping patterns to save passenger in-vehicle time is needed and is the focus of this study.

This paper has been organized as follows. In the next section, the characteristics and scale of the train stop planning problem are discussed. In the third section, an MIP model for optimizing the

---

J.-C. Jong, Civil, Hydraulic Engineering and Railway Transportation Research Center, Sinotech Consultant Engineering, Inc., No. 171, Nanking East Road, Section 5, Taipei 10570, Taiwan. C.-S. Suen and S. K. Chang, Division of Traffic Engineering, Department of Civil Engineering, National Taiwan University, Room 313, Civil Engineering Building, No. 1, Roosevelt Road, Section 4, Taipei 10617, Taiwan. Corresponding Author: J.-C. Jong, jcjong@sinotech.org.tw.

*Transportation Research Record: Journal of the Transportation Research Board*, No. 2289, Transportation Research Board of the National Academies, Washington, D.C., 2012, pp. 24–33.  
DOI: 10.3141/2289-04



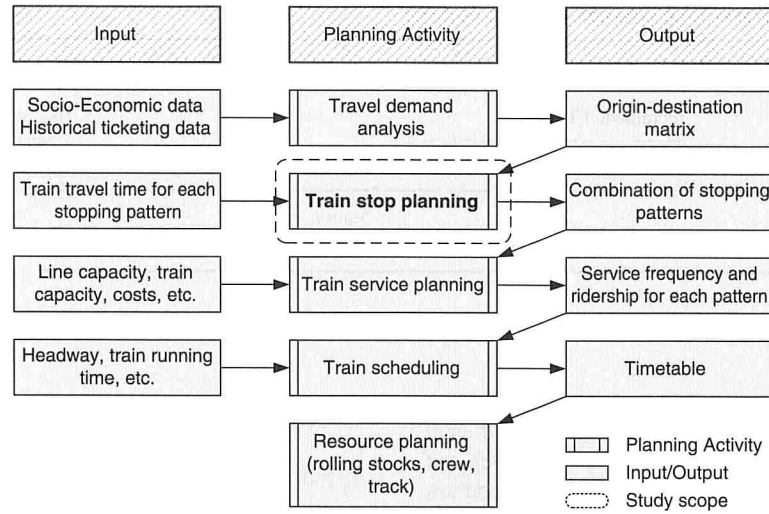


FIGURE 1 Planning process for a railway system.

combination of stopping patterns is formulated. Because the model cannot be efficiently solved by the traditional branch-and-bound method, in the following section a genetic algorithm (GA) is developed to accelerate the solution process. Finally, the Taiwan High Speed Rail (THSR) case is illustrated to demonstrate the efficiency and efficacy of the algorithm.

### TRAIN STOP PLANNING

Given the O-D matrix, the purpose of train stop planning is to generate an optimal set of stopping patterns. Because service frequencies and detailed costs are unavailable at this stage, the objective is usually to minimize passenger in-vehicle time. An all-stop train will meet all kinds of O-D pairs, but the in-vehicle time for some O-D pairs would be higher than that for patterns with fewer stops. However, express service would save passenger in-vehicle time for a particular O-D pair, but some O-D demands may not be served. Theoretically, if railway operators could provide direct services for each O-D pair, the total passenger in-vehicle time would be minimized. In reality, only a few stopping patterns (typically three to five) are provided because of limited resources and the complexity of railway operation. Thus, operators must determine an optimal combination of stopping patterns to save passenger in-vehicle time.

For a railway with  $n$  stations, there are  $n - 2$  intermediate stations through which trains may stop or pass, except for the origin and destination terminals. Therefore, the total number of stopping patterns, denoted by  $m$ , can be calculated as  $m = 2^{(n-2)}$ . If the operator decides to provide  $L$  types of stopping patterns for operations, then the problem scale is expressed as Equation 1:

$$C_L^m = \frac{m!}{L! \times (m - L)!} \quad (1)$$

For example, if a rail line has 10 stations and the operator wants to provide five types of stopping patterns, then there are more than 8.8 billion ( $C_5^{2^{(10-2)}} = 8,809,549,056$ ) possible combinations of stopping patterns. A greater number of stations  $n$  or expected patterns  $L$  result in a larger solution space. The problem scale for a real-world rail line may be tremendous and cannot be solved manually. Thus,

developing a rigorous approach to efficient selection of the optimal combination of stopping patterns is necessary.

### MIP FORMULATION

The train stop planning activity is inherently an optimization problem. The current study developed a MIP model that minimizes total passenger in-vehicle time, subject to practical constraints. The notation associated with the MIP model is listed in Table 1.

TABLE 1 Notation for MIP Model

Parameter	Description
$n$	Number of stations
$m$	Total number of stopping patterns, $m = 2^{(n-2)}$
$L$	Expected number of stopping patterns to operate
$P_{ij}$	Passenger demand from station $i$ to $j$
$a_{ij}^k$	Binary variable indicates whether pattern $k$ stops at stations $i$ and $j$ : If pattern $k$ stops at both stations $i$ and $j$ , then $a_{ij}^k = 1$ . If pattern $k$ does not stop at either station $i$ or $j$ , then $a_{ij}^k = 0$ .
$t_{ij}^k$	Travel time from station $i$ to $j$ for pattern $k$ . Calculation of $t_{ij}^k$ is illustrated in Tables 2 and 3.
$M_1, M_2$	Arbitrarily large positive constant, where $M_1 \ll M_2$
$s^k$	Binary variable to indicate whether pattern $k$ is selected for operation service: If pattern $k$ is selected for operation, then $s^k = 1$ . If pattern $k$ is not selected for operation, then $s^k = 0$ .
$T_{ij}^k$	Actual travel time from station $i$ to $j$ for pattern $k$ , depending on whether pattern $k$ is selected for operation ( $s^k$ ) and whether it stops at both stations $i$ and $j$ ( $a_{ij}^k$ )
$T_{ij}$	Minimal travel time from station $i$ to $j$ among all stopping patterns to be operated, i.e., $T_{ij} = \min(T_{ij}^1, T_{ij}^2, \dots, T_{ij}^k, \dots, T_{ij}^m)$
$y_{ij}^k$	Binary variable to specify whether the travel time from station $i$ to $j$ of pattern $k$ is the smallest among all patterns: If travel time from station $i$ to $j$ of pattern $k$ is the minimum, then $y_{ij}^k = 1$ . If travel time from station $i$ to $j$ of pattern $k$ is not the minimum, then $y_{ij}^k = 0$ .



TABLE 2 Example of Travel Time Between Adjacent Stations

Direction I (A → B)				Direction II (B → A)			
Station	Running Time	Incremental Time		Station	Running Time	Incremental Time	
		Deceleration	Acceleration			Deceleration	Acceleration
Station A	$t_1$	—	$t_3$	Station A	—	$t_5$	—
Station B	—	$t_2$	—	Station B	$t_4$	—	$t_6$

NOTE: — = not applicable.

Distinguishing train travel time between every O-D pair for each stopping pattern is important because of the existence of different stopping patterns. The method used in the Japan Shinkansen and THSR system is displayed in Table 2 and Table 3. The method can help users easily assess the travel time of an O-D pair for a particular stopping pattern. Table 2 is a typical example of travel time expression for both directions (e.g., downward and upward). Travel time between contiguous stations can be divided into several parts, namely, running time, acceleration time, deceleration time, and dwell time at stations. Table 3 shows that once the stopping pattern is decided, the travel time for that specific pattern can be easily calculated by simple addition.

The MIP formulation is illustrated below:

$$\min \sum_{i=1}^n \sum_{j=1}^n P_{ij} T_{ij} \quad (2)$$

subject to

$$T_{ij}^k = t_{ij}^k + M_1(1 - a_{ij}^k s^k) \quad \forall k, i, j \quad (3)$$

$$T_{ij} \geq T_{ij}^k - M_2(1 - y_{ij}^k) \quad \forall k, i, j \quad (4)$$

$$\sum_{k=1}^m y_{ij}^k = 1 \quad \forall i, j \quad (5)$$

$$\sum_{k=1}^m s^k = L \quad (6)$$

$$y_{ij}^k \leq s^k \quad \forall k, i, j \quad (7)$$

$$y_{ij}^k \in \{0, 1\} \quad \forall k, i, j \quad (8)$$

TABLE 3 Travel Time Between Adjacent Stations for Different Stopping Patterns

Station Scenario		Travel Time	
Station A	Station B	Direction I: Station A → Station B	Direction II: Station B → Station A
Nonstop	Nonstop	$t_1$	$t_4$
Nonstop	Stop	$t_1 + t_2$	$t_6 + t_4$
Stop	Nonstop	$t_3 + t_1$	$t_4 + t_5$
Stop	Stop	$t_3 + t_1 + t_2$	$t_6 + t_4 + t_5$

$$s^k \in \{0, 1\} \quad \forall k \quad (9)$$

$$T_{ij} \geq 0 \quad \forall i, j \quad (10)$$

$$T_{ij}^k \geq 0 \quad \forall k, i, j \quad (11)$$

The objective function (Equation 2) minimizes total passenger in-vehicle time, computed by multiplying passenger demand with the associated travel time. The solution process shall guide the search for suitable patterns to serve each O-D demand with minimal total in-vehicle time while satisfying constraints.

Equation 3 describes that if pattern  $k$  is selected for operation and stops at both stations  $i$  and  $j$  (i.e.,  $s^k = 1$  and  $a_{ij}^k = 1$ ), then the travel time from  $i$  to  $j$  of pattern  $k$  (i.e.,  $T_{ij}^k$ ) is just equal to  $t_{ij}^k$ . However, if pattern  $k$  is not selected for operation, or pattern  $k$  does not stop at either station  $i$  or  $j$  (i.e.,  $s^k = 0$  or  $a_{ij}^k = 0$ ), then the travel time from  $i$  to  $j$  of pattern  $k$  is set to an arbitrarily large positive constant  $t_{ij}^k + M_1$ . Equation 4 computes the minimal travel time from station  $i$  to  $j$  among all stopping patterns. In combination with Equation 3, if  $y_{ij}^k = 1$ , then  $T_{ij} \geq T_{ij}^k$ ; otherwise (i.e.,  $y_{ij}^k = 0$ ),  $T_{ij} \geq T_{ij}^k - M_2$ . In addition,  $T_{ij}^k - M_2 = t_{ij}^k + M_1 - M_2 < 0$  because  $T_{ij}^k \ll M_1 \ll M_2$  for all  $k$ . Thus,  $T_{ij} \geq T_{ij}^k - M_2$ , guaranteeing that Equation 4 will identify the minimal travel time from all selected patterns.

Equation 5 specifies that only one pattern could be selected to provide the minimal travel time for the specific O-D pair from  $i$  to  $j$ . Equation 6 specifies the summation of selected stopping patterns as  $L$ . This summation means that a trade-off exists between different patterns and only  $L$  patterns can be chosen for operations. Equation 7 confirms that the minimal travel time of a stopping pattern for a specific O-D pair is meaningful only when the pattern is selected. Equations 8 through 11 define some variables to be binary and others to be nonnegative.

The above model simultaneously determines whether a pattern is selected or not selected, as well as the minimal travel time for each O-D pair. The formulation is a MIP model with some binary variables. The model can be solved by the branch-and-bound or cutting plane methods, which are general algorithms for finding optimal solutions of MIP problems (18).

## SOLUTION APPROACH

The proposed MIP formulation can be used to solve small problems but is not practical for dealing with large-scale problems. To mitigate this situation and make application in a real system possible, the present study develops a GA model to handle the realistic problem. The advantages of using a GA model are global perspective, parallelism, and robustness. The model is simple, yet powerful in searching

TABLE 4 Notation for GA Model

Parameter	Description
$\mathbf{x}$	Binary vector to represent combination of stopping patterns in a chromosome
$x_k$	Binary variable as $k$ th gene of $\mathbf{x}$ to specify whether stopping pattern $k$ is selected for operation, where $1 \leq k \leq m$ : If pattern $k$ is selected, then $x_k = 1$ . If pattern $k$ is not selected, then $x_k = 0$ .
$\mathbf{z}^k$	Binary vector to represent stopping pattern $k$
$z_i^k$	Binary variable to specify whether pattern $k$ stops at station $i$ , where $1 \leq i \leq n$ : If pattern $k$ stops at station $i$ , then $z_i^k = 1$ . If pattern $k$ does not stop at station $i$ , then $z_i^k = 0$ .
$q$	Selective pressure in ordinal-based selection
$p_l$	Selection probability for $l$ th chromosome in ranking of population
$n_p$	Size of population

for improvement, and is not limited by assumed search space restrictions (19, 20). Specifications of the GA model are introduced below. Table 4 shows the associated notation. For convenience, a symbol  $U[b_l, b_u]$  is defined to denote an integer number from a discrete uniform distribution, whose domain is within  $b_l$  and  $b_u$ .

### Representation of Decision Variables

In a GA, the problem is treated as the environment, and a set of possible solutions to the problem is treated as the population. Each individual in the population is represented by an encoded solution called a chromosome. In this problem, a chromosome represents a combination of stopping patterns. The chromosome in the population is denoted by a binary vector as expressed in Equation 12.

$$\mathbf{x} = [x_1 \ x_2 \ \dots \ x_k \ \dots \ x_{m-1} \ x_m] \quad (12)$$

Here another binary vector  $\mathbf{z}^k$  is introduced as shown in Equation 13 to represent stopping pattern  $k$ , where  $z_i^k = 1$  denotes pattern  $k$  stops at station  $i$ ;  $z_i^k = 0$ , otherwise.

$$\mathbf{z}^k = [z_1^k \ z_2^k \ \dots \ z_i^k \ \dots \ z_{n-1}^k \ z_n^k] \quad (13)$$

Next, a mechanism must be established to relate  $x_k$  to its corresponding stopping pattern  $\mathbf{z}^k$ . If it is assumed that the first gene  $x_1$  in the chromosome represents the express (nonstop) service and the last gene  $x_m$  denotes the all-stop service, then the binary vector  $\mathbf{z}^k$  can be obtained from the representation of  $k - 1$  in the binary numeral system. For example, if a rail line has five stations, it will have  $2^3 = 8$  stopping patterns. The chromosome (i.e., the combination of stopping patterns) can be denoted by the binary vector  $\mathbf{x} = [x_1 \ x_2 \ x_3 \ x_4 \ x_5 \ x_6 \ x_7 \ x_8]$ . For each stopping pattern, both the first and last elements will be 1 because trains must stop at terminal stations (i.e.,  $z_1^k = z_5^k = 1$ ,  $\forall k = 1, \dots, 8$ ). The intermediate elements of the third stopping pattern are the binary representation of  $2 (3 - 1)$ , that is,  $2_{(10)} = 010_{(2)}$ . Similarly, the intermediate elements of the fourth stopping pattern are the binary representation of  $3 (4 - 1)$ , that is,  $3_{(10)} = 011_{(2)}$ . Table 5 summarizes the relations between chromosome, genes, and corresponding stopping patterns.

TABLE 5 Example of Chromosome, Genes, and Stopping Patterns

$x_k$ and $z_i^k$	Station 1	Station 2	Station 3	Station 4	Station 5
$x_1$	1	0	0	0	1
$x_2$	1	0	0	1	1
$x_3$	1	0	1	0	1
$x_4$	1	0	1	1	1
$x_5$	1	1	0	0	1
$x_6$	1	1	0	1	1
$x_7$	1	1	1	0	1
$x_8$	1	1	1	1	1

### Fitness Function

In evaluating chromosome fitness (i.e., the objective function), the minimal travel time for each O-D pair should be determined in advance. Once the gene values of a chromosome are given, the minimal travel time for each O-D pair  $i$  and  $j$  can be determined by

$$T_{ij} = \min\{T_{ij}^k\} \quad \forall s^k = z_i^k = z_j^k = 1 \quad (14)$$

With Equation 14 substituted into Equation 2, the chromosome fitness can be reformulated as follows:

$$f(\mathbf{x}) = \sum_{i=1}^n \sum_{j=1}^n P_{ij} \min\{T_{ij}^k\} \quad \forall s^k = z_i^k = z_j^k = 1 \quad (15)$$

### Initial Population

To maintain chromosome variability, the current study introduces some randomly generated solutions to the initial population. Real-world systems commonly use all-stop and express services, so these will be produced in the initial population to enhance the convergence characteristic of GA. The current study takes four types of solutions into the initial population. The first group contains an express service with other randomly selected patterns. The second group consists of all-stop services. The third group includes both express and all-stop services. The final solutions are randomly generated. The four types of initial solutions can be generated with the following procedures:

First group, including express service:

$$s^1 = 1$$

$$s^k = 0$$

$$\forall k = 2, \dots, m$$

Generate random number  $k \in U[2, m]$ , set  $s^k = 1$  until  $\sum_{k=1}^m s^k = L$ .

Second group, including all-stop service:

$$s^m = 1$$

$$s^k = 0$$

$$\forall k = 1, \dots, m - 1$$

Generate random number  $k \in U[1, m - 1]$ , set  $s^k = 1$  until  $\sum_{k=1}^m s^k = L$ .

Third group, including both express and all-stop services:

$$s^1 = s^m = 1$$

$$s^k = 0$$

$$\forall k = 2, \dots, m - 1$$

Generate random number  $k \in U[2, m - 1]$ , set  $s^k = 1$  until  $\sum_{k=1}^m s^k = L$ .

Fourth group, randomly generated:

Generate random number  $k \in U[1, m]$ , set  $s^k = 1$  until  $\sum_{k=1}^m s^k = L$ .

## Operators

The present research designs two types of genetic operators to guide the search, that is, mutation and crossover operators. The mutation operator explores the search space by arbitrarily changing some genes in the chromosome, whereas the crossover operator exploits information from parents by swapping their gene segments. The problem is a constrained optimization model, so the present work designs a repair procedure to legalize the solutions after application of genetic operators, such that they satisfy the constraints.

### Mutation Operator

To apply the mutation operator, a gene is randomly selected and its value is exchanged between 0 and 1. The total number of stopping patterns must be equal to  $L$ , so another gene must be chosen, and its value is exchanged accordingly. The procedures are illustrated as follows:

1. The amount of the selected gene (i.e.,  $x_k^l = 1$ ) in the chromosome is  $L$ , so a random number  $i \in [1, L]$  is generated to be the  $i$ th candidate gene in the selected rank.
2. The number of the remnant gene pool (i.e.,  $x_k^l = 0$ ) is  $m - L$ ; therefore, another random number  $j \in [1, m - L]$  is generated to be another  $j$ th candidate gene in the remnant rank.
3. The binary numbers of the  $i$ th and the  $j$ th are swapped.

### Crossover Operator

To apply the crossover operator for two randomly selected chromosomes, a position  $i$  between 1 and  $m$  is randomly generated, and then the gene segments of the two parents after  $i$  are swapped to create two offspring. The resulting offspring may not satisfy the constraints, so the current research develops a repairing mechanism to legalize the offspring. The procedures are summarized below:

1. Generate two random numbers  $a, b \in U[1, n_p]$  to select chromosomes for undergoing the crossover operator.
2. Generate a random number  $i \in U[1, m]$ .
3. Let  $\mathbf{x}^a = [x_1^a \ x_2^a \ \dots \ x_i^a \ \dots \ x_{m-1}^a \ x_m^a]$  and  $\mathbf{x}^b = [x_1^b \ x_2^b \ \dots \ x_i^b \ \dots \ x_{m-1}^b \ x_m^b]$  be the two chromosomes to be crossed after the  $i$ th gene. Then, the resulting offspring  $\mathbf{x}^{a'}$  and  $\mathbf{x}^{b'}$  are

$$\mathbf{x}^{a'} = [x_1^a \ x_2^a \ \dots \ x_i^a \ x_{i+1}^b \ \dots \ x_{m-1}^b \ x_m^b] \quad (16)$$

$$\mathbf{x}^{b'} = [x_1^b \ x_2^b \ \dots \ x_i^b \ x_{i+1}^a \ \dots \ x_{m-1}^a \ x_m^a] \quad (17)$$

4. For each offspring, if  $\sum_{k=1}^m x_k > L$ , the gene with a value equal to 1 is randomly selected, and its value is changed to 0 until  $\sum_{k=1}^m x_k = L$ . Similarly, if  $\sum_{k=1}^m x_k < L$ , the gene with a value equal to 0 is randomly selected, and its value is changed to 1 until  $\sum_{k=1}^m x_k = L$ .

## Reproduction and Replacement

In a GA model, a chromosome with better fitness has a greater chance of being selected for reproducing offspring. However, a selection based on raw fitness may result in premature convergence. The pro-

posed algorithm uses ordinal-based selection to prevent the population from sticking to local optima. The scheme selects individuals, not based on their raw fitness, but on their rank in the population. This method requires that selection pressure be independent of the population fitness distribution and based solely on the relative ordering of the population (21). Let  $q$  be a floating number representing the selective pressure that is input by the user between 0 and 1. Then, the selection probability  $p_l$  for the  $l$ th chromosome based on the ranking of the population is determined by the following equation:

$$p_l = \frac{q(1-q)^{l-1}}{1-(1-q)^{n_p}} \quad (18)$$

The current study introduces a reverse order to replace a worse chromosome. If a random number  $l$  is generated, then the actual chromosome selected to die is the  $(n_p - l + 1)$ th chromosome. Also, if  $l$  is equal to  $n_p$ , then the mechanism regenerates a random number because the best chromosome is never selected to die. This elitism model guarantees that the search is nondeteriorating.

## Genetic Algorithms Procedure

Figure 2 presents the basic structure of the proposed GA model for optimizing the combinations of train-stopping patterns.

Step 1. Parameters initialization. Station data, run time, and O-D demands are loaded into the system. Parameters for the GA model, such as population size, number of total generations, selective pressure, and termination rule, are inputted.

Step 2. Chromosome encoding. The decision variables are encoded as a binary vector, where 1 represents the corresponding stopping pattern selected; 0, otherwise.

Step 3. Population initialization. Four types of chromosomes are generated to initialize the population.

Step 4. Calculation of objective function value and updating. The fitness of each chromosome in the population is evaluated and updated by using Equation 15.

Step 5. Sorting of chromosomes according to their fitness. The chromosomes in the population are sorted by their fitness from small to large.

Step 6. Gene operation

–6.1: Selection. The elitist model is used to ensure that the evolution is nondeteriorating.

–6.2: Operator. Two types of genetic operators exist, namely, mutation and crossover. Each type is performed at certain times, as determined by input parameters.

Step 7. Repairing chromosomes. Each offspring is ensured to meet the constraints of the expected number of stopping patterns.

Step 8. Steps 4–7 are repeated until the population converges, and then Step 9 is executed.

Step 9. The best solution is outputted.

## CASE STUDY

The current study uses the C++/MFC technique, one of the most popular programming languages with Microsoft Foundation Class library, to pre-proceed with the parameters, O-D matrix, and MIP formulations. The system then uses the optimization solver CPLEX to derive the optimal solution with the branch-and-bound method and produces the output.

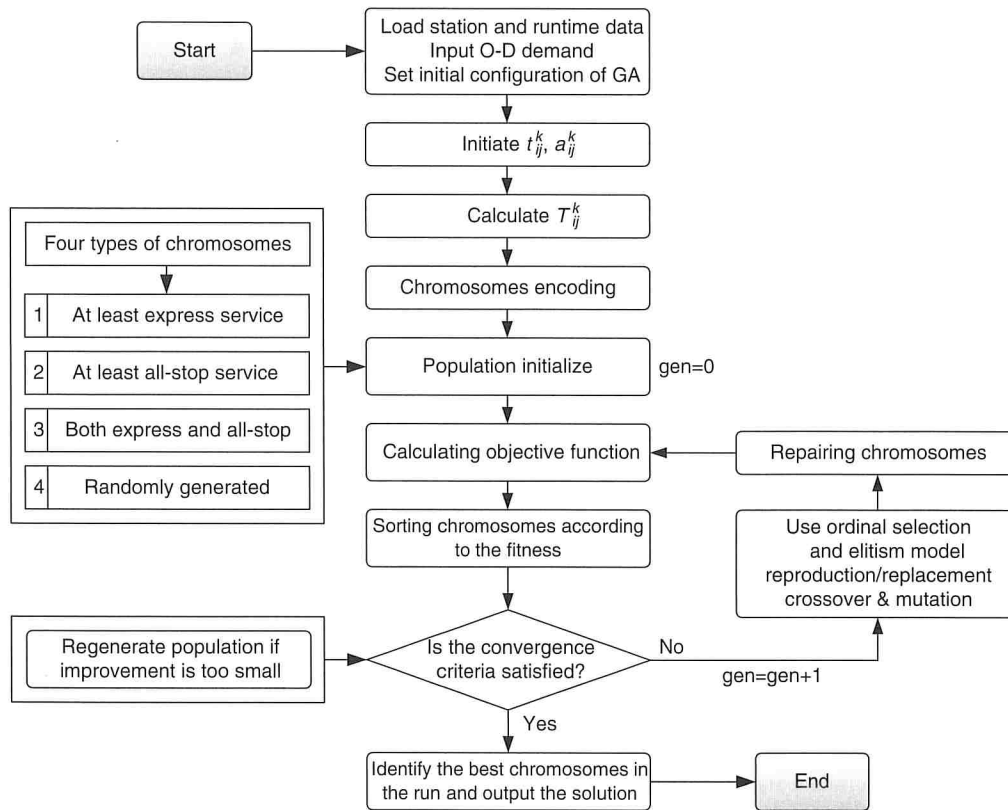


FIGURE 2 GA flowchart for optimizing train-stopping patterns.

The GA model is also coded in the C++/MFC programming technique in combination with several other techniques to accelerate convergence. In addition, the present study introduces two approaches to confirm the convergence characteristic of the GA model. One approach is to calculate for a sufficient number of times by using different initial populations. If all results are the same, then the algorithm is believed to be in convergence. Another approach is to compare the results with those of the MIP model.

The current study demonstrates the performance of the proposed MIP and GA models. A desktop PC (Core 2 Duo CPU, 3.0 GHz with 4 GB RAM) is used to execute the proposed algorithm. The program verification meets the requirements of the study. The minimal objective value and optimal set of stopping patterns are identical between the MIP and GA models, thereby validating the robustness of the GA model. Figure 3a shows an example of providing two or three stopping patterns under the situation of four to eight stations. Figure 3a shows that while the station number is greater than six, the solution time of the MIP model increases exponentially. The problem scale for providing three patterns under eight stations is so huge that the MIP model could not converge even with a 3-day computation time. Figure 3b shows that the GA model performs more efficiently under the same conditions. Therefore, this section proceeds to exhibit the applicability of the GA model in minimizing the in-vehicle time for all passengers.

### Taiwan High Speed Rail System

The THSR case study and the existing condition of the study are briefly described. Figure 4 shows the scope of the current study, in which the THSR Corporation (THSRC) possesses eight stations at the initial stage. The THSR system is approximately 345 km in length

along the west corridor of Taiwan, and the average daily ridership was more than 100,000 passengers in 2010. The rail connects three major cities, namely, Taipei, Taichung, and Zouying, whose stations can be terminals in the system. Five stopping patterns (5A\*–5E\*) are specified from the initial planning scheme. However, these patterns were politically and empirically determined by the Bureau of Taiwan High Speed Rail, which means that the patterns may be incapable of meeting most travel demands. In other words, improvements can still be made.

In practice, the real operation situation shows that the THSRC adopts mainly patterns 5B\* and 5D\* for daily services and uses the 5C\* and 5E\* patterns only to balance the train between the terminal and depot at dawn or midnight. Finally, Pattern 5A\* is not used all along. As mentioned previously, operators provide few stopping patterns because of the resource limitations. The proposed model is then applied to observe the relationship between total passenger in-vehicle time and the number of stopping patterns (Figure 5).

Figure 5a shows the objective values of the optimal solution according to the expected number of stopping patterns to operate. Each optimal solution under a specific  $L$  has the minimal value among all different combinations of stopping patterns. In regard to  $L$ , a greater number of stopping patterns result in less total passenger in-vehicle time. In addition, Figure 5b shows that the average generations of convergence increase substantially when the problem scale increases. A good fine-tuning measure for the GA parameters, such as increasing the number of generations, mutations, or crossovers, may improve the situation.

### Optimal Combination of Stopping Patterns

The current study follows the actual limitations of the THSRC and takes Taipei and Zouying as the normal termini of the studied route.

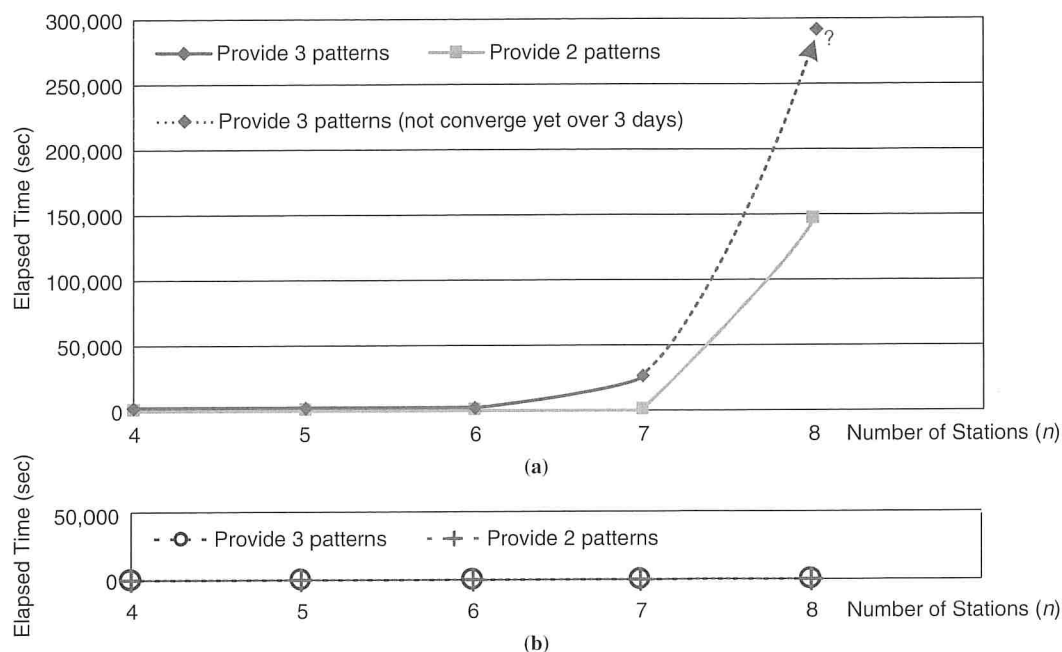


FIGURE 3 MIP and GA execution elapsed time versus scale of problem for (a) MIP model and (b) GA model.

Taichung is not an appropriate terminal station because of depot design and environmental regulations.

Figure 6 demonstrates several scenarios for different combinations of stopping patterns in accordance with the actual O-D demand of the THSRC. Although an all-stop service results in longer travel time, it guarantees that passenger demands can be served by any O-D pair, which is why an all-stop service is the fundamental type of service in most scenarios. However, if the THSRC adopts sufficient types of stopping patterns (e.g., seven patterns in Figure 6f), a number of other service patterns can be substituted for all-stop to gain the optimal result.

From a geographical viewpoint, a short distance exists between Taipei and Banciao Station. Therefore, the THSRC regards Banciao as a hinterland of Taipei and allows every train to stop at this location.

In contrast, results from the current study show that Banciao is no longer a must-stop station. For example, two major existing patterns used by the THSRC are 5B\* and 5D\*. Supposing that only two patterns are provided, Figure 6a suggests that to meet travel demands, the original pattern 5B\* should stop at Tainan instead of Banciao. In comparison, Figure 5a also shows that the optimal solution can save 16.5 h (i.e., 59,494 s, down from 6,527,529 to 6,468,035 s) a day, whereby 6,032 h a year could be saved for passengers in this situation. Moreover, the proposed model produces more significant benefits, saving 112.2 h a day, compared with the existing four patterns (excluding 5A\*).

In the case of direct services between two end stations, such as Taipei–Taichung, Taichung–Zuoying, Taipei–Hsinchu, and Taipei–Zuoying, visibilities are progressively introduced into the results.

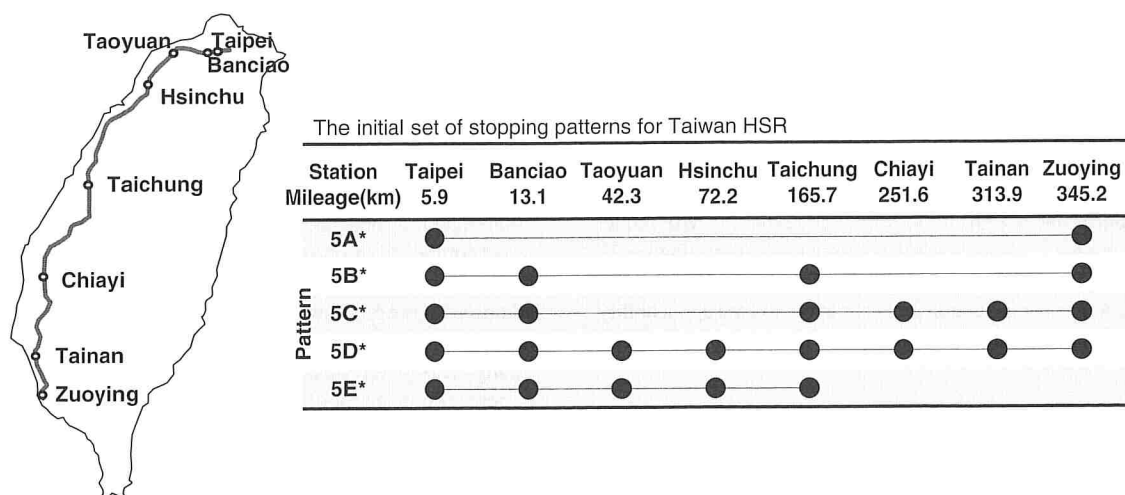


FIGURE 4 Rail line and original planned patterns of THSR (adapted from Bureau of High Speed Rail, Taiwan).

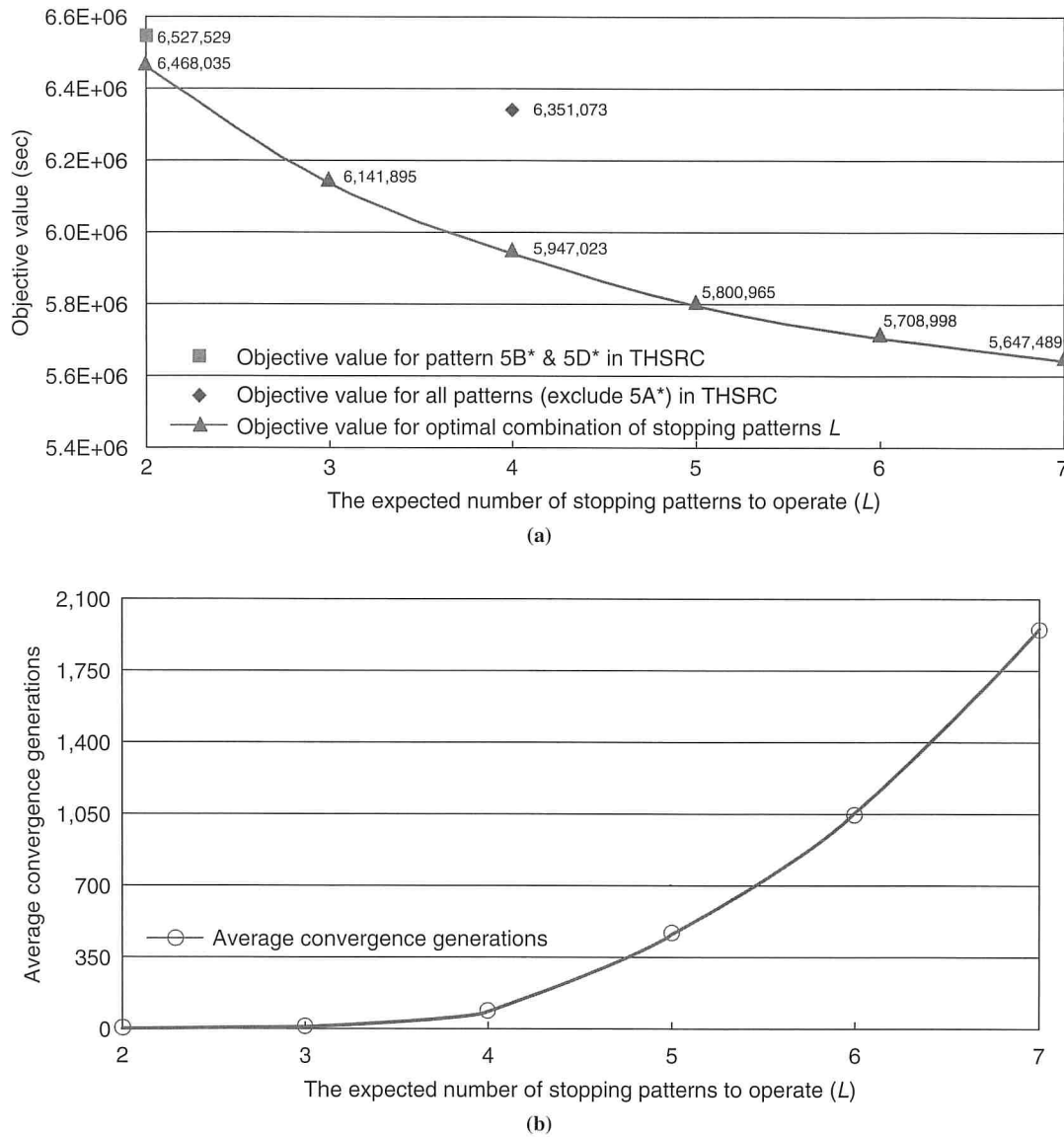


FIGURE 5 Objective value and generations versus number of stopping patterns: (a) objective value versus stopping patterns  $L$  and (b) average convergence generations versus stopping patterns  $L$ .

For all scenarios, the optimal sets of stopping patterns are varied and based on the O-D pairs. Hence, the THSRC should review these patterns periodically to ensure that all service patterns meet actual demands. Moreover, the opening of four additional stations in the near future is a challenge for the corporation. Should the THSRC provide four service patterns under 12 stations, more than 45.5 billion ( $C_4^{(12-2)} = 45,545,029,376$ ) solution combinations will be possible. The proposed decision support system and optimization process can help operators make decisions efficiently and correctly.

## CONCLUSIONS

Simultaneously considering all components in the service planning process in a model and solving a full-range problem is difficult because of the heterogeneous demand and the complexity of an intercity passenger rail. To improve this planning process, the GA

model is developed to carry out the optimal combination of stopping patterns in accordance with limitations and to minimize total passenger in-vehicle time.

The robustness of the GA model and its efficient implementation are validated by the MIP model. Convergence generation increases in large-scale problems, but the model still performs well through the fine-tuning of GA parameters.

Results show that this well-developed support system can assist railway operators in planning their stopping patterns correctly and efficiently. The method can also be helpful for a newly introduced system, such as those proposed high-speed rail systems in the United States, especially when no historical data are available for reference. With the resulting stopping patterns, planners could build train service planning models such as those developed to determine the optimal service frequency for each pattern (10, 11). The required number of train sets can also be estimated by the round-trip time and optimal service frequency of each stopping pattern.



Pattern	Taipei	Banciao	Taoyuan	Hsinchu	Taichung	Chiayi	Tainan	Zuoying
2A	●				●		●	●
2B	●	●	●	●	●	●	●	●

(a)

Pattern	Taipei	Banciao	Taoyuan	Hsinchu	Taichung	Chiayi	Tainan	Zuoying
3A	●				●			●
3B	●	●				●	●	●
3C	●	●	●	●	●	●	●	●

(b)

Pattern	Taipei	Banciao	Taoyuan	Hsinchu	Taichung	Chiayi	Tainan	Zuoying
4A	●				●			●
4B	●			●			●	●
4C	●	●	●			●		●
4D	●	●	●	●	●	●	●	●

(c)

Pattern	Taipei	Banciao	Taoyuan	Hsinchu	Taichung	Chiayi	Tainan	Zuoying
5A	●							●
5B	●				●	●		●
5C	●			●			●	●
5D	●	●	●		●			●
5E	●	●	●	●	●	●	●	●

(d)

Pattern	Taipei	Banciao	Taoyuan	Hsinchu	Taichung	Chiayi	Tainan	Zuoying
6A	●							●
6B	●				●			●
6C	●			●			●	●
6D	●	●				●		●
6E	●	●	●				●	●
6F	●	●	●	●	●	●	●	●

(e)

Pattern	Taipei	Banciao	Taoyuan	Hsinchu	Taichung	Chiayi	Tainan	Zuoying
7A	●							●
7B	●						●	●
7C	●				●	●		●
7D	●			●	●			●
7E	●	●	●					●
7F	●	●	●		●		●	●
7G	●	●	●	●		●	●	●

(f)

FIGURE 6 THSRC scenarios for optimal combination of stopping patterns: (a) two stopping patterns, (b) three stopping patterns, (c) four stopping patterns, (d) five stopping patterns, (e) six stopping patterns, and (f) seven stopping patterns.

## REFERENCES

1. Assad, A. A. Modelling of Rail Networks: Toward a Routing/Makeup Model. *Transportation Research 14B*, 1980, pp. 101–114.
2. Bussieck, M. R., T. Winter, and U. T. Zimmermann. Discrete Optimization in Public Rail Transport. *Mathematical Programming*, Vol. 79, 1997, pp. 415–444.
3. Sussman, J. *Introduction to Transportation Systems*. Artech House ITS Library, 2000.
4. Landex, A., A. H. Kaas, and S. Hansen. *Railway Operation*. Publication Report 2006-4. Centre for Traffic and Transport, Technical University of Denmark, Kongens Lyngby, 2006.
5. Sone, S. Novel Train Stopping Patterns for High-Frequency, High-Speed Train Scheduling. In *Computers in Railways III, Technology*, Vol. 2 (T. K. S. Murthy, J. Allan, R. J. Hill, G. Sciutto, and S. Sone, eds.), Computational Mechanics Publications, United Kingdom, 1992, pp. 107–118.
6. Eisele, D. O. Application of Zone Theory to a Suburban Rail Transit Network. *Traffic Quarterly*, Vol. 22, 1968, pp. 49–67.
7. Deng, L. B., F. Shi, and W. L. Zhou. Stop Schedule Optimum of Passenger Train Plan. *Sciencepaper Online*. June 2008. [www.paper.edu.cn](http://www.paper.edu.cn). Accessed March 1, 2009.
8. Hung, C. Y. *Train Plan Model for Taiwan High Speed Rail*. MS thesis. National Cheng Kung University, Taiwan, 1998.
9. Chen, Y. Y. *Integration of High Speed Rail System Scheduling and Train Seats Allocation*. MS thesis. National Taiwan University, Taiwan, 1998.
10. Jong, J. C., and C. S. Suen. A Train Service Planning Model with Dynamic Demand for Intercity Railway Systems. *Proc., Eastern Asia Society for Transportation Studies*, Vol. 6, 2007, pp. 1598–1613.
11. Jong, J. C., and C. S. Suen. Optimizing Train Service Plan for Intercity Railways. *Proc., 8th World Congress on Railway Research (CD-ROM)*, 2008.
12. Hsieh, W. J. Service Design Model of Passenger Railway with Elastic Train Demand. *Proc., Eastern Asia Society for Transportation Studies*, Vol. 5, 2003, pp. 307–322.
13. Lee, C. K., and W. J. Hsieh. A Bilevel Programming Model for Planning High Speed Rail Service. *Transportation Planning Journal*, Vol. 31, No. 1, 2002, pp. 95–119.
14. Nichols, F. High-Speed Train Service Plan—Full Build Network with Links to Sacramento and San Diego. Technical memorandum (draft), 2009.
15. Albrecht, A. R., and P. G. Howlett. Application of Origin–Destination Matrices to the Design of Train Services. *Australasian Journal of Engineering Education*, Vol. 15, No. 2, 2009, pp. 95–104.
16. Chang, Y. H., C. H. Yeh, and C. C. Shen. A Multiobjective Model for Passenger Train Services Planning: Application to Taiwan's High-Speed Rail Line. *Transportation Research*, Vol. 34B, 2000, pp. 91–106.
17. Ulusoy, Y. Y., S. I.-J. Chien, and C.-H. Wei. Optimal All-Stop, Short-Turn, and Express Transit Services Under Heterogeneous Demand. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 2197, Transportation Research Board of the National Academies, Washington, D.C., 2010, pp. 8–18.
18. Rardin, R. L. *Optimization in Operations Research*. Prentice Hall, Upper Saddle River, N.J., 1998.
19. Yin, Y. F. Genetic-Algorithm-Based Approach for Bilevel Programming Models. *Journal of Transportation Engineering*, Vol. 126, No. 2, 2000, pp. 115–120.
20. Michalewicz, Z. *Genetic Algorithm + Data Structure = Evolutionary Programs*, 3rd ed. Springer-Verlag, New York, 1996.
21. Miller, B. L., and D. E. Goldberg. Genetic Algorithms Selection Schemes and the Varying Effects of Noise. *Evolutionary Computation*, Vol. 4, No. 2, 1996, pp. 113–131.

---

*The Intercity Passenger Rail Committee peer-reviewed this paper.*

# Risk Assessment of Positive Train Control by Using Simulation of Rare Events

Timothy Meyers, Amine Stambouli, Karen McClure, and Daniel Brod

The risk assessment of positive train control (PTC) presents a number of challenges that can be addressed through simulation, a common tool for analyzing large, complex stochastic systems. The combined analysis of a simulated rail system with safety models that track the propagation of human errors and equipment failures toward hazards and accidents (or their eventual safe resolution) enables the prediction of accidents and their probability of occurrence for a base case without PTC and an alternate case with PTC. Accidents are rare events, and when probabilities of rare events are estimated, efficiency is a major concern because the computer resources required for statistically reliable estimates are usually overwhelming. The problem of efficiency can be addressed through multilevel splitting, or staged simulation. The basic idea of splitting is to create separate copies of the simulation whenever it approaches the rare event. The FRA generalized train movement simulator (GTMS) integrates a rail system simulator with safety models and staged simulation to arrive at metrics of safety and risk that meet federal regulatory requirements. The simulation techniques used and a description of their implementation in the GTMS are presented. The paper concludes with a case study risk assessment that uses the GTMS of a nonvital overlay PTC system for a Class I railroad.

For a number of years FRA has sought to develop tools that support positive train control (PTC) risk assessment and the evaluation of system safety risk in general, by focusing efforts on simulation methods. These methods derive from simulation of the railroad physical plant, human agents, and the causal chains leading to accidents provided by FRA and the railroad industry. As such, the methods hold the promise of improved transparency in developing simulated results that describe risk while confirming likely sequencing of hazardous events leading to accidents and other unsafe incidents.

Conventional simulation methods, such as Monte Carlo, are adequate for examining the probabilities associated with common operational occurrences. However, using Monte Carlo simulation to derive statistically reliable estimates of rare events, such as accidents and their predecessor events, requires enormous computational resources, thus rendering Monte Carlo simulation impractical for this purpose.

---

T. Meyers, A. Stambouli, and D. Brod, DecisionTek, LLC, 6337 Executive Boulevard, Rockville, MD 20852. K. McClure, Federal Railroad Administration, Office of Railroad Policy and Development, 1200 New Jersey Avenue, SE, Mail Stop 20, Washington, D.C. 20590. Corresponding author: T. Meyers, [tmeyers@decisiontek.com](mailto:tmeyers@decisiontek.com).

*Transportation Research Record: Journal of the Transportation Research Board*, No. 2289, Transportation Research Board of the National Academies, Washington, D.C., 2012, pp. 34–41.  
DOI: 10.3141/2289-05

In its development of a generalized train movement simulator (GTMS) system, FRA has implemented an alternative to Monte Carlo simulation. The alternative uses staged simulation, which generates the sought-after analytic outcomes while using a fraction of the computer resources. This paper describes the methods that were implemented in the GTMS and presents the findings of a PTC risk assessment for a nonvital overlay system that was conducted with the GTMS.

## RESEARCH BACKGROUND

The Rail Safety Improvement Act of 2008 mandates the implementation and operation of PTC systems by 2015 on intercity passenger rail lines and on Class I freight railroads in which annual tonnage exceeds a specified threshold. PTC provides added safety, at a minimum, through in-cab information and enforced braking, which stops trains at the end of their movement authority and prevents overspeeding and work zone incursions.

A nonvital PTC overlay system is a type of PTC system that operates in conjunction with the existing traffic control and braking systems. In the event of a PTC system failure, the existing systems remain fully operational and perform their safety critical functions. According to the 2010 PTC Rule, the initial installation of nonvital overlay systems requires a risk assessment as part of the approval process. The risk assessment must demonstrate that the level of risk on a rail network using a proposed PTC system is 80% lower than the risk before PTC installation. Subsequent PTC installations or modifications on the same rail network must maintain at least the same level of safety as the initial PTC installation.

The 80% reduction in risk for PTC preventable accidents can be expressed by forecast accidents and incidents, or measures of frequency (i.e., mean time to accident or accidents per million train miles), that reflect the inherent safety risk in the system. Risk assessment results are presented for a base case scenario (before proposed PTC implementation) and an alternate case scenario (after proposed PTC implementation), which permits a side-by-side comparison and straightforward evaluation of the reduction in risk provided by the new technology.

The challenge of conducting PTC risk assessments lies in the ability to provide results that reflect the accumulated risk of the proposed system over its life cycle while ensuring that these results are statistically reliable. Several approaches exist for assessing rare-event probabilities. In railroads, the events leading to hazards and the evolution of a hazard to an incident or accident, along with the severity of the accident, are highly dependent on the operating environment and the specifics of each accident. From the point of view of event specificity, simulation techniques are perhaps the best means for assessing safety risk. Researchers have recognized for

some time that simulation techniques hold significant promise for PTC risk assessment. Several studies were conducted on the topic in the late 1980s and early 1990s by the Draper Laboratory and in the early 2000s by the University of Virginia's Center of Railroad Safety—Critical Excellence.

With Monte Carlo simulation, the computer system simulates railroad operations for an extended period of time and generates accidents or other hazardous events of interest to arrive at statistically reliable estimates of the frequency of accident occurrence. However, generating a sufficient number of rare events through simulation for statistically reliable estimates will overwhelm any computing system and take an unacceptably long time to produce useful results.

A more appropriate methodology for rare-event simulation is the use of a multilevel splitting technique, which splits the simulation into stages (or levels) at each of a series of events known to lead to an accident. Each time an event occurs that brings the simulation closer to the sought-after rare event, the system state is stored. These stored system states are used as starting points for the next simulation level. In this way, the problem space is reduced and the analysis focuses on those paths that have some probability of culminating in an event of interest while ignoring those paths that have no such probability. This technique yields a comprehensive risk assessment that is conducted within practical constraints while providing statistically reliable outcomes.

## THEORETICAL APPROACH

In simulation, a rare event probability  $\gamma$  is estimated by dividing the number of observed occurrences by the number of trials  $n$  (in this context a "trial" can be viewed as an hour of railroad operations). A measure of statistical reliability is the relative error, defined as the standard deviation of the estimator divided by its mean. The standard error of the estimator  $\gamma$  is approximately  $1/\sqrt{n\gamma}$ ; therefore, the number of required trials grows inversely with the square of the desired relative error.

An alternative approach that has been successfully applied to rare event problems is splitting, or staged simulation. The premise of staged simulation is to create copies of the simulation state at each split or occurrence of an event that brings the system closer to the rare event of interest ( $I$ ). When sufficient splits are collected, they are used as the starting points for the next simulation stage. By defining multiple stages (or levels) at which a split can occur, the staged simulation technique preserves simulation resources, by focusing only on generating the events that have a better chance of leading to the sought-after rare event. The accuracy of this method depends on how the splits are defined and the number of events harvested at each level.

Staged simulation is well suited for predicting railroad accidents or incidents. Generally, the path to a train accident or incident is forged by a well-known sequence of events, or causal chain, which incrementally elevates the risk of the system until all preconditions for an accident are met. For example, one possible causal chain for a head-to-head collision accident occurs as follows: (a) a train crew fails to initiate on-time braking when approaching its end of authority; (b) the train exceeds its authority, by entering a block in which it has no authority to proceed; and (c) a second train is granted authority for the block it enters and may collide with the first train depending on their relative positions and speeds. Each event in this example brings the system closer to an accident and is thus defined as the start of a new simulation level.

Staged simulations are conducted in levels. In each level, all available computing resources are used to generate events of interest, or occurrences, for that level. In the first level, Level 0, the sought-after events are those that initiate the causal chains that lead to accidents. With the previous example, a Level 0 event would be "train crew fails to initiate on-time braking when approaching its end of authority." During a Level 0 simulation, trains are permitted to run in the system for a specified time period (say, 5 years). When a Level 0 event occurs, the simulation does the following: (a) the system state is captured and stored and then (b) the human error or system failure is corrected for continued safe rail operations. The "system state" is the entire simulated railroad operating environment at the time of the occurrence and includes the time, position, and speed of each train; the position of each switch; the aspect of each signal; and all movement authorities that have been granted by the central dispatcher and traffic control system. At the end of a Level 0 simulation, a pool of system states has been captured at each point where a causal chain originating event has occurred.

The next simulation level, Level 1, seeks to generate the events that extend the causal chains initiated in Level 0. Revisiting the previous example, a Level 1 event would be "the train exceeds its authority, entering a block in which it has no authority to proceed." To generate Level 1 events, the Level 1 simulation randomly samples from the pool of system states captured in Level 0 and resumes each simulation at the point at which its system state was stored. By simulating in this manner, each simulation trial begins from a Level 0 event and has a better chance of generating a Level 1 event, bringing the system closer to generating the rare event. When a Level 1 event occurs, the simulation does the following: (a) the system state is captured and stored and then (b) the simulation trial ends and thus prompts the Level 1 simulation to sample a new system state from the Level 0 pool.

A staged simulation can have as many levels as needed to control the unfolding of causal chains. All simulation levels after Level 1 follow the same process, sampling from the previous level's pool of system states to generate a new event of interest in the causal chain. In the final level, rare events are generated. With the previous example of head-to-head collisions, the probability of such accidents can be estimated after a sufficient number of them are generated by using a series of outputs produced in each level of the staged simulation.

The probability of a head-to-head collision can be stated as the mean time to accident, defined as

$$MTT A_{HHC} = \frac{MTTH}{P_{HHC|EAH}} \quad (1)$$

where MTTH is the mean time to hazard, or the Level 1 event of interest from which the accident was generated. The variable  $P_{HHC|EAH}$  is the probability of a head-to-head collision, given that a hazardous condition has occurred. In this case, a train exceeds its authority and encroaches on the authorized path of another train.

At each level, the probability of the level event, or  $p$ , is equal to the number of occurrences divided by the number of simulation trials required to generate those occurrences. The conditional probability of a rare event rail accident is  $p_1 * p_2 * \dots * p_n$ , where  $n$  is the number of simulation levels.

The mean time at each level event is the mean time to the previous level event divided by the current level probability, except for Level 0. The mean time to the Level 0 event of interest is equal to the total hours of Level 0 computer run time divided by the number

of errors or failures generated during that time. The formulas for staged simulation metrics are given below.

### Level 0

Mean time to error or failure is defined as

$$MTTE = \frac{T_E}{N_E} \quad (2)$$

where  $T_E$  is the total hours of operations in Level 0 and  $N_E$  is the number of error and failure events generated in Level 0.

### Level 1

The probability of a hazardous event, given a human error or equipment failure, is defined as

$$P_{H|E} = \frac{N_H}{n_{T1}} \quad (3)$$

where  $N_H$  is the number of hazardous events generated in Level 1 and  $n_{T1}$  is the number of Level 1 trials.

Mean time to hazardous event is defined as

$$MTTH = \frac{MTTE}{P_{H|E}} \quad (4)$$

### Level 2

The imputed probability of an accident is defined as

$$P_{A|H} = \frac{N_A}{n_{T2}} \quad (5)$$

where  $N_A$  is the number of accidents generated in Level 2 and Level 1 and  $n_{T2}$  is the number of Level 2 trials.

Mean time to accident is defined as

$$MTTA = \frac{MTTH}{P_{A|H}} \quad (6)$$

Conditions for the sufficiency of the duration of the Level 0 simulation and the number of trials for Levels 1 and 2 simulations, as well as an optimal allocation of computer resources across levels, can be derived [see Shortle et al. (1)]. A simple test for the sufficiency of the number of trials at each level is that the estimated mean conditional probability and its variance are stable and do not change with added trials.

## IMPLEMENTATION IN GTMS

The theoretic approach described above was implemented in FRA's GTMS software. The GTMS contains a train movement model and a train dispatcher model, which are overlaid with a risk assessment model that generates accidents and other rare events of interest. The train movement model calculates the forces on the train, including

the tractive effort, the braking force, and the resistive force. The dispatcher model determines the path of trains through the simulated rail network and grants authorities for movement.

The GTMS uses a hybrid of fixed time interval and discrete event simulation in which train movements are calculated as discrete events and are synchronized to fixed time intervals (usually 60 to 180 s, but can be reduced to as little as 5 s to capture a very granular sequence of events when unsafe situations occur).

### Train Movement Model

The GTMS train movement model replicates realistic train movements over a rail system calculating the forces acting on the train while considering terrain (grade), track geometry (curvature), track speed limits, and the specific consists of simulated trains. As a train moves along its route, the changing forces on the train determine its position, time, and speed in the simulated system (2).

The train receives its routing information and authority to move from the dispatcher model, and the train accelerates and decelerates according to its effective speed limit, which is derived from the track speed limit, granted authorities, and any speed restrictions in effect. Given the train consist—the list of locomotives and cars that make up the train—and the track, the trains advance with small incremental changes in speed until the forces on the train are in balance (subject to the speed limit). The resistive force on the train is recalculated on a car-by-car basis every 500 ft to account for changes in grade and track curvature. The train movement model determines a preferred throttle position in accordance with best train handling practice, which determines the tractive effort for the specified locomotive consist. The braking algorithm simulates dynamic braking with partial service air braking and full service air braking as the last choice (or as an “enforced” option in the event of PTC corrective action).

### Dispatcher Model

The dispatcher model operates on a node network that is overlaid on the real-world network of control blocks. A node represents an area of the simulated rail network that can be authorized only to a single train at a time. The dispatcher model determines the path of trains through the network and grants authorities for movement (3). Authorities are granted so as to achieve safe separation of trains and facilitate train meets and the overtaking of lower priority trains. The dispatcher grants an authority to a train only if the movement of the train is free of conflict and will prevent deadlock. Authorities are revoked only after a train has executed a movement authority and exited the authorized block. Through the dispatcher model and the configuration of control blocks, alternative train control systems can be simulated. The dispatcher model lends itself to parameterization and implementation of traffic control strategies that replicate traffic control of real-world alternatives, such as direct traffic control or centralized traffic control (CTC).

### Safety Model and Causal Chains

The GTMS risk assessment begins by defining the causal chains that link hazards, accidents, and incidents. Human errors and equipment failures initiate the causal chains that evolve into hazards and accidents or incidents or resolve safely, depending on the interaction of trains, train crews, and dispatchers.



GTMS causal chains were developed in close cooperation with the FRA Office of Safety and the Class 1 railroads through the Railroad Safety Advisory Committee process established by FRA. Each causal chain begins with either human errors or equipment failures. The initiating human errors occur when train crews fail to observe operational directives from the dispatcher model. These human errors include failures to initiate on-time braking, heed work zones, and heed impending speed restrictions. Initiating equipment failures occur when simulated switches are misaligned or set against movement authority. Next, each causal chain links the human or equipment response to the initiating errors and failures. For example, train crews or the PTC system can intervene with corrective braking measures in response to the initial human errors.

After the initiation of a human error or equipment failure event, the interaction of trains, train crews, and dispatchers can allow one or more of the following hazards to occur:

- End of authority hazard occurs when a train enters track for which it has no movement authority.
- Misaligned switch hazard occurs when a train intersects a switch that is set in neither the normal nor the reverse position (misaligned).
- Unauthorized switch alignment hazard occurs when a train intersects a switch that is aligned against proper movement authority.
- Work zone incursion hazard occurs when a train encroaches into a work zone.
- Overspeed hazard occurs when a train crew violates an approaching speed restriction.

Finally, the hazards described above can evolve into one of the following accidents or incidents:

- Overspeed derailment,
- Emergency braking derailment,
- Enforcement braking derailment,
- Work zone accident or incident,
- Unauthorized alignment switch derailment,
- Misaligned switch derailment,
- Head-to-head collision,
- Head-to-tail collision, and
- Sideswipe collision.

## GTMS Model Verification and Validation

The GTMS software developers, FRA staff, and participating Class 1 railroads have conducted GTMS model runs and reviewed the results to confirm that the train movement and dispatcher models successfully replicate railroad operations for the traffic levels, signaling systems, and railroad networks under review.

## CASE STUDY

### GTMS Risk Assessment

The GTMS risk assessment of a proposed nonvital PTC overlay system deployed by a Class I railroad finds that the system would eliminate up to 95% or more of the accident risk compared with the existing CTC system. The risk assessment was conducted in accordance with the requirements of the PTC Rule as set forth in Appen-

dix B to 49 CFR Part 236. Appendix B outlines the risk assessment criteria for systems that fall under Subpart H of the rule, which includes nonvital overlay PTC systems. The appendix describes the risk metrics, risk computation principles, and major systems and subsystems whose risks are to be included in a risk assessment.

The proposed system enforces compliance with the existing underlying CTC system, operating rules, and procedures and provides added protection against the consequences of human error and equipment failure. Railroad systemwide component failure rates and human error probabilities were used for the analysis.

Base case accident rates were within 5% of railroad industry rates on similar territory and operating environments. Alternative case results were reviewed by the railroad officials and the FRA Office of Safety and found acceptable. Alternative case results also aligned with the PTC component system failure rates.

## Territory of Study and Operational Scenario

### *Territory Description*

The rail system under study is 160 mi long and has very mild grades—usually less than 0.5%. The territory runs from northwest to southeast and has an interchange with another Class I railroad. The territory is mostly single-track with passing sidings. The numerous sidings in the territory have not been upgraded to handle trains with a 286,000 loaded car weight and will not carry trains in excess of 12,000 tons. This weight limit restricts traffic of loaded unit and coal trains to the main track, and when trains meet, the lower-weight train is always diverted to the siding. The overall track speed limit for the territory is 49 mph. Movements through switches in the reverse direction are restricted to 10 or 20 mph.

### *Traffic Control*

The base case traffic control in the territory is CTC. In a CTC system, opposing and following train movements are authorized and governed by block signals, and the signal indication is the source of authority for the train crew. A proceed signal provides the needed train movement authority.

### *Operational Scenario*

The high-traffic scenario assumes that 54 coal, unit, and general merchandise trains per week traverse the territory. The simulated trains operate continuously every day during the simulation period. Daily traffic varies from five to 12 trains per day and is made up mainly of empty and loaded coal trains that are up to 135 cars and 7,300 ft long. Loaded cars weigh up to 286,000 lb. The simulated period of operations is 25 years, which is the assumed life span of the proposed PTC system. During this period of analysis, 70,435 trains traveling 10,285,865 mi were simulated over the subdivision.

### *Description of PTC System*

The nonvital PTC system in the alternate case interfaces with the existing or base case CTC traffic control systems. The purpose of the PTC system is to mitigate the effects of potentially hazardous operational errors or equipment failures by enforcing compliance



with train movement authorities, speed restrictions, switch positions, and work zones and includes the following functionality:

- Movement authority enforcement
  - Predictively enforces end of authority with 75 s of a visual alert accompanied at the start by a momentary audible alert before enforcement,
  - Reactively protects against revoked authorities, and
  - Includes protection at territory entrance, transition, and exit (predictive on unambiguous track, reactive on ambiguous track);
- Speed limit enforcement
  - Pertains to all permanent and temporary speed limits,
  - Predictively enforces impending reduced speed limits with 75 s of a visual alert accompanied at the start by a momentary audible alert before enforcement, and
  - Reactively enforces overspeed condition while providing audible and visual alerts (no specific duration) after overspeed occurs until enforcement threshold is reached;
- Work zone enforcement
  - Predictively enforces entrance into unacknowledged work zone with 75 s of a visual alert accompanied at the start by a momentary audible alert before enforcement and
  - Reactively enforces continued movement after stopping in a work zone and
- Wayside detection
  - Includes misaligned switch detection and broken rail detection and is provisioned for landslide detection, high-water detection, high-wind detection, high- and wide-load detection, misaligned bridge detection, warning bearing notification, dragging equipment notification, failed highway crossing, and other special signal devices.

## Risk Assessment Inputs

The risk assessment inputs are used to populate the safety model and the staged simulation framework. Safety model parameters include human error rates, equipment failure rates, and accident or incident probabilities. These parameters intervene at different levels of the staged simulation described in the section on risk assessment methodology.

Human errors and equipment failures occur during Level 0 of the staged simulation and determine whether hazardous events are generated in Level 1. Accident and incident probabilities determine whether hazardous events generated in Level 1 resolve safely or result in an accident or incident in Level 2 of the staged simulation.

### Human Errors and Equipment Failures

**Rate of Train Operator Error** The GTMS Safety Model relies on well-established human factors models and research to estimate the probability of human error, defined as the number of errors committed per 1,000 h of train operations (4). In the simulation model, a train operator commits an error in one of three ways:

- Train operator fails to deploy conventional braking on approaching the train's end of authority (in accordance with the simulation model train handling conventions, a full stop is implemented with dynamic braking combined with partial service air brakes with 10 psi set),

- Train operator fails to heed impending speed restriction, and
- Train operator fails to heed an impending work zone.

Given the rate of error and the train operator unreliability for a shift  $t_0$  hours long, the probability of error when action is required (an exponentially distributed random variable) is given by the formula

$$F(t) = 1 - e^{-\beta t_0} \quad (7)$$

where  $\beta$  is the rate of operator error and  $t$  is the length of operator shift in hours. The analysis assumes an operator shift of 10 h. Each time a train approaches its end of authority, a speed restriction, or a work zone, when the operator is required to brake (or heed an impending speed restriction or work zone), a random number is generated on (0, 1) (the interval of real numbers between 0 and 1) and if the value is less than that given by the above formula, then the simulation model triggers a human error event.

### Given Train Operator Error, Mean Time Until Corrective Action Taken

In the event that a train operator commits a fail to heed end of authority error, the simulation model predicts the time elapsed (in seconds) until the operator realizes his or her error and initiates corrective action (i.e., applies emergency brakes). The time elapsed is modeled as an exponentially distributed random variable, calculated by using the following formula:

$$F(t) = 1 - e^{-\frac{t}{\mu}} \quad (8)$$

where  $\mu$  is the mean time to corrective action (in seconds) and  $t$  is the time elapsed since the occurrence of the human error. The mean time to corrective action  $\mu$  is set by using the "given train operator error, mean time until corrective action taken" parameter.

### Probability of Misaligned Switch Given Approaching Train

In the event that a train approaches a switch, the simulation model uses the "probability of misaligned switch given approaching train" parameter to predict whether the approaching switch is in a misaligned state. A misaligned switch is one that is set in neither the normal nor the reverse position. If the switch is misaligned, the train will not be given authority to proceed (the analysis assumes zero probability of failing to detect a misaligned switch).

### Probability That Switch Is Aligned Against Movement Authority Given Approaching Train

In the event that a train approaches a switch, the simulation model uses the "probability that switch is aligned against movement authority given approaching train" parameter to predict whether the approaching switch is set in an unauthorized position. A misaligned switch is one that is set in neither the normal nor the reverse position. If the switch is found to be set in the wrong position, the train will not be given authority to proceed (the analysis assumes zero probability of failing to detect a switch aligned against movement authority).

**Rate of PTC Failure to Warn (Failures per Hour)** In Alternate Case Risk Assessments (i.e., simulations of PTC-enabled rail systems), a warning is issued to the train crew in the event that

- The train operator fails to brake on approaching the train's end of authority,

- The train operator fails to heed an impending speed restriction, or
- The train operator fails to heed an approaching work zone.

The parameter “rate of PTC failure to warn” is an exponentially distributed random variable that determines whether the PTC system fails to operate correctly and warn the train crew to take action and avoid an unsafe condition. If the PTC equipment fails to warn the train crew, then it will attempt to enforce braking if the train crew fails to take corrective measures.

**Rate of PTC Failure to Enforce Braking (Failures per Hour)** In alternate case risk assessments, PTC enforces braking when

- The train crew fails to acknowledge PTC’s warning of an impending hazard or
- PTC fails to warn the train crew of an impending hazard.

GTMS uses the “rate of PTC failure to enforce braking” as the parameter of an exponentially distributed random variable to determine whether the PTC equipment will enforce braking and stop the train before a hazard occurs. If the PTC equipment fails to enforce braking, the train crew may still correct and attempt to manually stop the train. If the crew fails to brake, then a hazard will occur.

#### *Probability of Accident or Incident Given Hazardous Situation Parameters*

**Probability of Derailment from Emergency Braking** Given a train operator error, the simulation model calculates the time elapsed until corrective action is initiated (i.e., deployment of emergency brakes). When emergency brakes are applied, the simulation model uses the “probability of derailment from emergency braking” parameter to determine whether the brake application results in a derailment.

**Probability of Derailment for Misaligned Switch or Unauthorized Switch Alignment** When a train approaches a switch that is misaligned or aligned against authorized movement, the signaling system detects the equipment failure and displays a restrictive aspect. If the train operator fails to heed the signal, the train can intersect the switch. The simulation model uses the “probability of derailment for misaligned switch or unauthorized switch alignment” parameter to predict whether the train’s intersection with the switch results in a derailment.

**Probability of Derailment Given Overspeed Hazard** When a train operator fails to heed an impending speed restriction, he or she can produce an overspeed hazard. The simulation model uses the “probability of derailment given overspeed” parameter to predict whether the overspeed results in a derailment.

**Probability of Accident or Incident Given a Work Zone Incursion** When a train operator fails to heed an approaching work zone, a work zone incursion hazard can result. The simulation model uses the “probability of accident or incident given a work zone incursion” parameter to predict whether the incursion results in a work zone accident or incident.

**Probability of Derailment from Enforcement Braking** When PTC enforces braking, the simulation model uses the “probability

of derailment from enforcement braking” parameter to determine whether the enforcement braking results in a derailment.

**Probabilities of Derailments Given Hazards** The probabilities of derailment given hazards were derived from published studies and expert opinion.

#### *Safety Model Parameter Values*

The parameters for the case study are shown in Table 1. Many of the safety model inputs were derived from industry averages, published studies, and expert opinion, and others were based on empirical or experiential-based information.

#### *Accident Severity*

Average accident severity costs per accident are based on publicly reported railroad data. Each accident type was assigned to one of two severity categories. The cost per incident in Category 1 (“less severe”) was \$168,837 and for Category 2 (“more severe”) the cost per accident was \$1,829,542. These per accident costs were derived from the Class I railroad’s average cost per accident for the 10-year period 1986 to 2005.

More severe accidents included head-to-head collisions, head-to-tail collisions, and sideswipe collisions. The less severe accidents included emergency braking and enforced braking derailments, misaligned switch derailment, work zone accident or incident, and overspeed derailment.

#### *Staged Simulation Parameters*

At each level of the simulation (0, 1, and 2), simulation parameters were selected to control simulation duration and randomness. Random seeds were used to yield a unique sequence of pseudorandom numbers to populate random variables at each level.

**TABLE 1 Risk Assessment Parameter Values**

Safety Model Parameter	Error or Failure Rate
Rate of train operator error (errors/h)	0.0004
Given train operator error, mean time until corrective action taken (s)	20
Probability of misaligned switch given approaching train	0.01
Probability that switch is aligned against movement authority given approaching train	0.005
Probability of derailment from emergency braking	0.05
Rate of PTC failure to warn (failures/hr of train operations)	0.005
Rate of PTC failure to enforce (failures/hr of train operations)	0.005
Probability of derailment from enforcement braking	0.005
Probability of derailment for misaligned or mis-set switch	0.05
Probability of derailment given overspeed	0.005
Probability of accident or incident given a work zone incursion	0.01

For the Level 0 simulation, the required values are the start and end dates for the simulated period of operations. In Levels 1 and 2, the required values are the number of trials.

For the period of operations set in Level 0, the simulation allows for the occurrence of human errors and equipment failures. When an error or failure event occurs, the GTMS stores the system state at the point of occurrence for reuse as randomly sampled initial conditions in Level 1 trials. After storing the system state, the GTMS rolls back the failure or error and continues to operate safely until the next error or failure occurs.

In Levels 1 and 2 the number of trials determines the number of times that previous-level stored system states are drawn at random, reanimated, and simulated until a hazardous occurrence or a safe resolution. The resulting probability of the level of interest event is then calculated as the number of occurrences encountered for the selected number of trials.

### Risk Assessment Results

This section presents risk assessment results for a high-traffic scenario (54 trains per week).

#### Level 0 Results

The Level 0 analysis simulation produced 990 Level 0 events during 25 years of simulated operations for both design cases. Results in Level 0 for the base and alternate cases are identical because PTC does not prevent errors and failures from occurring. Table 2 displays the Level 0 simulation results and mean time to event by error and equipment failure type.

#### Level 1 Results

The Level 1 analysis simulation produced 9,195 Level 1 events in the base case and 345 Level 1 events in the alternate case during 25 years of simulated operations. Table 3 shows the Level 1 risk assess-

TABLE 2 Level 0 Events by Type of Error and Failure

Level 0 Event	Scenario	Number of Level 0 Events in 25 years	Mean Time to Level 0 Event (MTTE) (days)
Fail to heed work zone	Base case (non-PTC)	70	130.43
	Alternate case (with PTC)	70	130.43
Fail to brake	Base case (non-PTC)	225	40.58
	Alternate case (with PTC)	225	40.58
Fail to heed speed restriction	Base case (non-PTC)	695	13.14
	Alternate case (with PTC)	695	13.14

ment results, which include the number of Level 0 human errors or equipment failures that were sampled (trials), the number of hazards generated when all sampled simulations were resumed, and the mean time to hazard implied by the results. It is observed from these results that PTC prevented hazards for a majority of unsafe conditions originating from human errors or equipment failures.

#### Level 2 Results

The Level 2 analysis simulation produced 435 Level 2 events in the base case and 170 Level 2 events in the alternate case during 25 years of simulated operations. Table 4 shows the Level 2 risk assessment results. The number of Level 1 hazards sampled (trials), the number of accidents generated when all sampled simulations were resumed, and the mean time to accident implied by the simulation outcomes are presented. It is observed from these results that the nonvital PTC overlay system prevented accidents of all types stemming from hazardous situations initiated by human errors or equipment failures. The proposed nonvital PTC overlay system significantly reduced the overall rail system operational risk.

TABLE 3 Level 1 Risk Assessment Results

Level 1 Event	Scenario	Level 0 Event	Number of Level 0 Event Trials	Number of Level 1 Events	Probability of Level 1 Event Given a Level 0 Event	Mean Time to Level 1 Event (MTTH) (days)
Work zone hazard	Base case (non-PTC)	Fail to heed work zone	735.00	735	1	130.43
	Alternate case (with PTC)	Fail to heed work zone	795.00	50	.06	2,073.79
End of authority hazard	Base case (non-PTC)	Fail to brake	1,975.00	1,275	.646	62.85
	Alternate case (with PTC)	Fail to brake	2,405.00	15	.0062	6,505.9
Overspeed hazard	Base case (non-PTC)	Fail to heed speed restriction	6,995.00	6,995	1	13.14
	Alternate case (with PTC)	Fail to heed speed restriction	6,800.00	290	.043	308.03
Exceeded authority hazard (misaligned switch)	Base case (non-PTC)	Fail to brake	135.00	115	.852	47.63
	Alternate case (with PTC)	Fail to brake	140.00	0	0	More than 300 years
Exceeded authority hazard (unauthorized switch alignment)	Base case (non-PTC)	Fail to brake	160.00	140	.875	46.37
	Alternate case (with PTC)	Fail to brake	125.00	0	0	More than 300 years

TABLE 4 Level 2 Risk Assessment Results

Level 2 Event	Scenario	Level 1 Event	Number of Level 1 Event Trials	Number of Level 2 Events	Probability of Level 2 Event Given Level 1 Event	Mean Time to Level 2 Event (MTTA) (days)
Work zone accident	Base case (non-PTC)	Work zone hazard	830	15	.018	7,216.98
	Alternate case (with PTC)	Work zone hazard	1,415	0	0	Over 300 years
Head-to-head collision	Base case (non-PTC)	End of authority hazard	1,650	65	.039	1,596.04
	Alternate case (with PTC)	End of authority hazard	550	0	0	Over 300 years
Head-to-tail collision	Base case (non-PTC)	End of authority hazard	1,650	0	0	Over 300 years
	Alternate case (with PTC)	End of authority hazard	550	0	0	Over 300 years
Sideswipe collision	Base case (non-PTC)	End of authority hazard	1,650	250	.152	414.97
	Alternate case (with PTC)	End of authority hazard	550	25	.045	143,130
Emergency brake derailment	Base case (non-PTC)	End of authority hazard	1,435	65	.039	1,387.65
	Alternate case (with PTC)	End of authority hazard	550	10	.018	357,825
Overspeed derailment	Base case (non-PTC)	Overspeed hazard	7,445	15	.002	6,520.12
	Alternate case (with PTC)	Overspeed hazard	8,585	135	.016	19,588.42
Misaligned switch derailment	Base case (non-PTC)	Exceeded authority hazard (unauthorized switch alignment)	100	10	.1	463.7
	Alternate case (with PTC)	Exceeded authority hazard (unauthorized switch alignment)	0	0	0	Over 300 years
Unauthorized switch derailment	Base case (non-PTC)	Exceeded authority hazard (misaligned switch)	115	15	.1304	482
	Alternate case (with PTC)	Exceeded authority hazard (misaligned switch)	0	0	0	Over 300 years

### Increased Human Error Due to Complacency

One of the features of nonvital overlay PTC systems is that the underlying safety critical systems remain in effect. This means that should PTC fail in part or in total, train operators and dispatchers will have pre-PTC capabilities at their disposal. In the event of a system failure, train crews will be able to ascertain all locational and directive information through means that were available in the base case (i.e., written instructions and signals) and will be able to bring the train to a stop manually.

One of the issues to consider is whether the presence of a non-vital overlay system leads to a sense of complacency. Because train operators will know that PTC warning and enforced braking will, under normal circumstances, automatically deploy in the event of an unsafe condition, some believe that operators will develop a sense of complacency. Complacency, should it occur, will be manifest in a higher operator error rate.

As part of the analysis of sensitivity, the case study ran the alternate (with PTC) case while assuming that the error rate was 25% higher, that is, 0.0005 error per operating hour in comparison with the previous assumption of 0.0004 error per operating hour. All other parameters of the analysis were left unchanged.

The analysis finds that an increase of 25% in operator error resulted in a 121% increase in safety-related costs. This result, however, still far exceeds the 80% risk reduction threshold that is required by the PTC Rule (i.e., \$5.3 million/million train miles

needs to drop to a level of \$1.06 million/million train miles, whereas PTC with increased complacency reduces risk to \$16,900/million train miles).

### ACKNOWLEDGMENTS

The authors thank the following former and current FRA personnel: Magdy El-Sibaie and Sam Alibrahim for support and guidance in developing the GTMS, Olga Cataldi for her valuable input, and Bor-Chung Chen for his review of the simulation methodology and statistical formulas.

### REFERENCES

1. Shortle, J. F., C.-H. Chen, B. Crain, A. Brodsky, and D. Brod. Optimal Splitting for Rare Event Simulation. *IIE Transactions*, 2010.
2. *USDOT/TSC Train Performance Simulator (TPS) User's Manual*, Version 5. Transportation Systems Center, U.S. Department of Transportation, 1988.
3. Lu, Q., M. Dessouky, and R. C. Leachman. Modeling Train Movements Through Complex Rail Networks. *ACM Transactions on Modeling and Computer Simulation*, Vol. 14, No. 1, 2004, pp. 48–75.
4. Dhillon, B. S. *Human Reliability and Error in Transportation Systems*. Springer, London, 2007.

*The Railroad Operating Technologies Committee peer-reviewed this paper.*

# Dual-Mode and New Diesel Locomotive Developments

Janis Vitins

The integration of electric and diesel traction into a single rail vehicle is technically challenging because of weight and space restrictions, particularly for AC catenary power. Through the combination of recent developments in power converter technology, diesel engine design, and mechanical lightweight structures, dual-mode locomotives are now feasible for railroad applications. Apart from the different modes of traction, such locomotives must also fulfill the latest vehicle standards in regard to safety, environmental impact, and interoperability. A key component is the DC link of the traction converter, which interfaces to electric and diesel power supply systems. In addition, batteries, supercaps, or both can be interfaced to this DC link. All electric power flow is bidirectional and thus permits many possibilities for energy savings and reductions in exhaust emissions. The performance of a dual-mode locomotive is greatly enhanced by the latest high-speed diesel engines developed for off-road and industrial applications. Not only do the engines provide high diesel power at low weight, but they also meet the new Tier 3 and upcoming Tier 4 exhaust emission standards. For maximum vehicle performance in both modes, the car body and truck must be lightweight. A monocoque car body and fabricated truck are the obvious solutions, as used on the dual-powered ALP-45DP and the European TRAXX AC3, which is an electric locomotive with a diesel engine for operation on nonelectrified sidings and terminals. The above technologies also lend themselves to new diesel-electric locomotives, by yielding a high vehicle performance at low axle loads, as required for passenger services at 125 mph.

Passenger trains have been the traditional domain of locomotive traction. Given the necessity for new and extended passenger services, economic solutions are needed for new types of locomotives that can run beyond the existing service networks. Whereas diesel traction is widespread in North America, there is also an electrified network, primarily in the Northeast, as well as third-rail systems in many larger cities and suburban areas. Further electrification can be expected as ridership increases in the future and thus the need to haul longer and heavier trains in high-density traffic. In this environment, the dual-mode locomotive can play an important role, allowing a one-seat ride over system borders, both electric and diesel. This is currently the situation at the New Jersey Transit Corporation (NJ Transit) and Agence Métropolitaine de Transport (AMT), Montreal. The ALP-45DP is presently in delivery for these railroads (1). These locomotives have become technically feasible in past years thanks to the advent of compact and lightweight propulsion

equipment and by combining this equipment with the latest diesel engine developments. Manufacturers of off-road and industrial diesel engines have invested heavily in new engine technologies, thus making these machines attractive also for railroad applications. These engines open up new opportunities for fuel savings and for lowering exhaust emissions in compliance with the upcoming Tier 4 requirements.

## PROPULSION CONCEPT OF DUAL-MODE LOCOMOTIVES

In a dual-mode locomotive, the traction converter must interface with the electric power source from the catenary, the third rail, or both, as well as with the power output from the alternator. This interface is best accomplished with the voltage-source traction converter, which has an intermediate DC link (see Figure 1). This DC link is composed of a capacitor bank to which power input and output circuitry are connected. Under catenary, the line voltage is reduced by means of the traction transformer to levels compatible with the converter switching devices. The output from the secondary windings is then rectified and fed into the DC link. In third-rail applications, the power can flow directly through an input choke or via a step-up chopper into the DC link. In diesel traction, the power output of the alternator is rectified and also fed into the same DC link. Therefore, the DC link is the common interface to the traction motors and drive systems. The DC link voltage is determined by the rated blocking voltage of the insulated gate bipolar transistor (IGBT) switching devices of the converter. For 3.3 kV IGBT devices, the DC link voltage is typically 1,700 V, and for 4.5 kV IGBTs it is approximately 2,800 V. The choice of IGBT device, and thus the DC link voltage, depends on many design factors. In general, a high DC link voltage is chosen for high traction power.

The DC link can be viewed as the common bus bar for power distribution in the locomotive. In the case of passenger locomotives, additional inverters can be connected to this DC link to feed the passenger cars with head-end power (HEP). Also, the locomotive auxiliary converters draw power from the DC link, and batteries and other electric storage devices can be added in the same way. In all cases, modern propulsion technology allows power to flow in both directions. This flow pattern allows the following possibilities of power flow in traction and dynamic braking:

- Traction. Power flow from overhead catenary or third rail to traction motors, auxiliaries, HEP, and energy storage devices;
- Traction. Power flow from diesel engine to traction motors, auxiliaries, HEP, and energy storage devices and, if desired, also back to the catenary; in the latter case, diesel engine full load testing is possible without resistor grids for power dissipation;

---

Bombardier Transportation, Brown-Boveri-Strasse 5, Zurich CH-8050, Switzerland.  
janis.vitins@ch.transport.bombardier.com.

*Transportation Research Record: Journal of the Transportation Research Board*, No. 2289, Transportation Research Board of the National Academies, Washington, D.C., 2012, pp. 42–46.  
DOI: 10.3141/2289-06



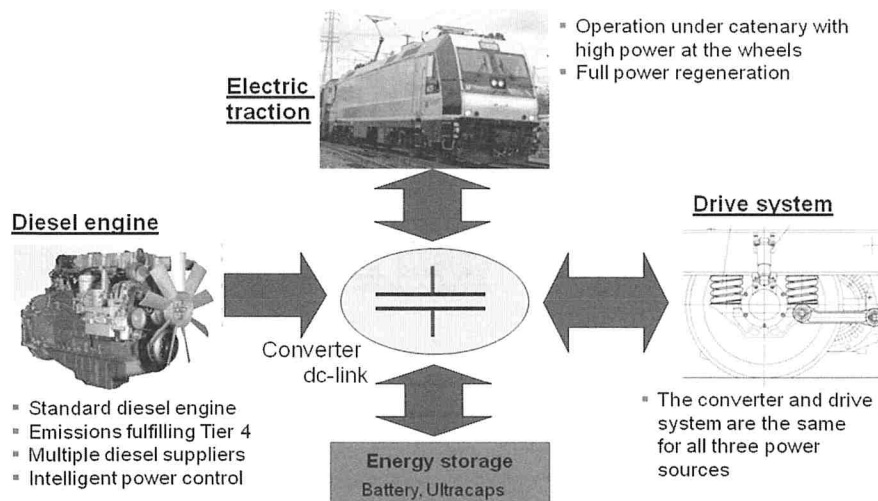


FIGURE 1 DC link of traction converter is common interface for electric and diesel power flow; in addition, energy storage devices can interface to DC link.

- Traction. Power flow from energy storage devices to traction motors, HEP, auxiliaries, and catenary; and
- Dynamic braking. Power flow from traction motors to catenary (or third rail), HEP, auxiliaries, and energy storage; therefore, even in diesel mode there is a nonnegligible amount of power regeneration.

The power rating of each power source depends on the specific application. Typically, a high power is needed in the electric mode for fast acceleration of long, high-capacity trains to high speeds. Nonelectrified networks are generally less congested and often have lower maximum speeds. Therefore, the power at the wheels from the diesel engines can usually be somewhat less than in electric traction. Energy storage devices today are still not capable of providing economically sufficient traction power for reasonable traveling distances. Therefore, their application is limited mostly to smaller shunters and special purpose vehicles. However, this limitation may soon change with the availability of new battery technologies.

### ALP-45DP DUAL-POWERED LOCOMOTIVE

The ALP-45DP dual-powered locomotive (see Figure 2) was developed initially for NJ Transit and the Montreal-based railroad AMT to operate their passenger trains on routes that are only partly equipped with overhead catenary. The objective is to provide passengers with a one-seat ride on services with mixed diesel and electric traction. The dual-powered locomotive also enables railroads to switch off diesel traction under catenary, in instances in which diesel locomotives have previously been used. The ALP-45DP thus allows substantial fuel savings, with diesel traction being substantially more expensive overall than electric traction under catenary.

The challenge in the design of the ALP-45DP was to provide sufficient power ratings in both electric and diesel mode under the constraint of a maximum axle load of 72,000 lb (32.66 metric tons) and meeting the required FRA Standards 49 CFR 229 and 238. This was done by using the latest technologies already incorporated in the electric locomotive ALP-46A for the car body, trucks, propul-

sion, and control and communication. Also, the operator's cab is basically identical to that of the ALP-46A, with additional functions for diesel traction. The fuel tank has a capacity of 1,800 U.S. gal and is an integral part of the car body structure. The tank consists of four separate compartments of 450 U.S. gal each to fulfill the stringent requirements of fire safety in tunnels. The trucks are derived from the German locomotive BR 101 and are of the same basic design as is used on the ALP-46 locomotives but adapted to the higher axle load.

The major engineering challenge was to install both the diesel engines and the electric traction equipment with the space and weight restrictions of a four-axle locomotive. Two Caterpillar 3512HD high-speed engines were chosen for the diesel mode. One advantage of these engines is that they can provide high acceleration, essential to achieve short traveling times between stations. Also, the engines meet Tier 3 exhaust emission standards. A solution is now being prepared to meet Tier 4. The disposition of equipment in the locomotive is shown in Figure 3. The traction converter is positioned in the center of the locomotive, with the transformer directly beneath it. This arrangement has the advantage of short power cables between them. The two diesel engines are mounted left and right of the converter, providing a largely symmetric layout and thus facilitating the



FIGURE 2 ALP-45DP for AMT (left) and NJ Transit (right).



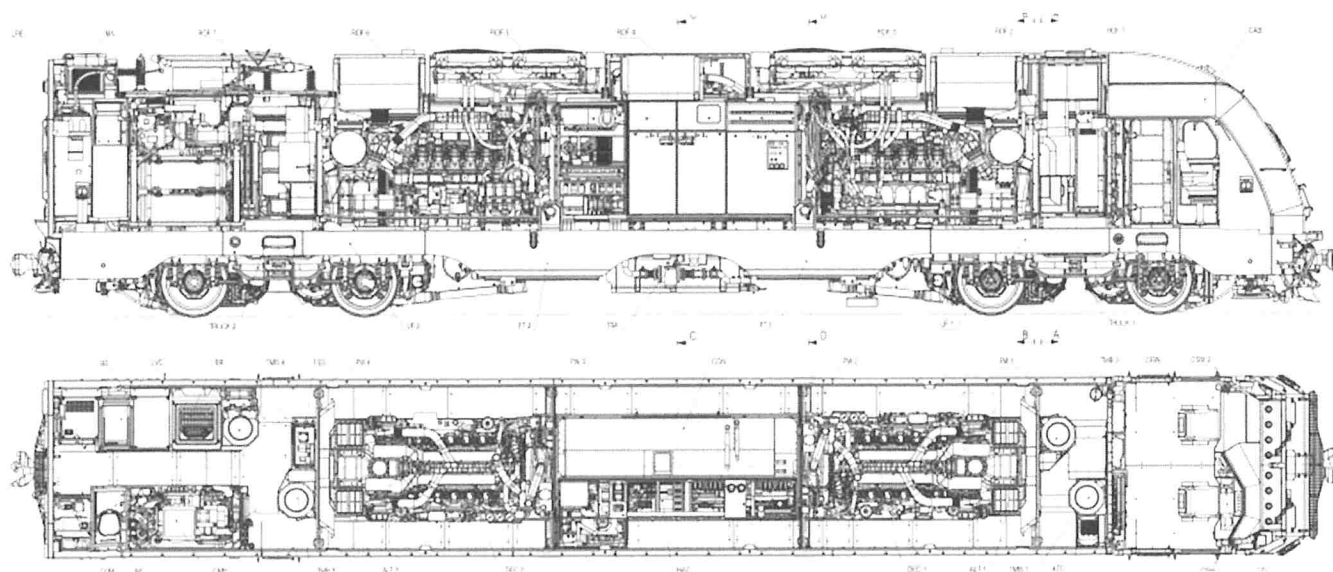


FIGURE 3 Equipment disposition in ALP-45DP.

weight balance of the locomotive. The cooling systems are installed in the roof sections directly above the diesel engines, thus making the best use of available space. The fuel tanks are directly below the diesel engines. The rheostatic brake cubicle, the cabinets for low voltage and the auxiliary drive distributor, battery boxes, and the toilet are mounted in the rear section of the locomotive.

The different energy flows from the catenary and diesel engines are handled by the same traction converter and drive system (2). In the electric mode, the catenary voltages used on the Northeast Corridor of 25, 12.5, or 12 kV (60 and 25 Hz) are reduced in the transformer and are rectified for the DC link by the line converters. In diesel mode, the line converters take a new function and rectify the output voltage of the alternators and feed DC power into the traction converters. The diesel engines are turned on by the alternators, with power sourced from the line converters and batteries. The key technical data of the ALP-45DP are shown in Table 1. The ALP-45DP locomotive has a low environmental impact in relation to noise and exhaust emissions, described as follows:

- The diesel engines can be turned off in the stationary mode under overhead lines. The locomotive is then parked with the pantograph up, such that electric power from the catenary can be used to heat or cool the train. This feature is important in densely populated areas.
- When the catenary is not available, noise and exhaust emissions can be significantly reduced through the use of only one diesel engine.
- The locomotive can be operated with just one diesel engine at slow speeds and at low power demand.

The above measures lead to a considerable drop in fuel consumption compared with conventional diesel locomotives, and this drop, in turn, permits the use of somewhat smaller fuel tanks. Because the total costs (operating and maintenance) are significantly higher for diesel than for electric operation, the dual-mode locomotive also results in considerable overall cost savings to the railroads. The ALP-45DP also improves operational availability. If one diesel

engine fails, the locomotive still has full tractive effort and can continue operation with half power. The HEP supply is maintained and the journey continued.

As shown above, the propulsion concept of the dual-mode locomotive also allows for the integration of a third rail capability for services, for example, in the New York region. The important additions needed for the ALP-45DP are the pickup shoes and high current cabling and switches, as well as input and chopper chokes. Because of the large currents drawn from the third rail, such chokes add several tons of weight and require adequate space for mounting. First investigations show that it is possible to include third rail with the ALP-45DP, however, only with smaller and lighter diesel engines so as to remain within the axle load limits and within the existing length of the locomotive.

TABLE 1 Key Technical Data for the ALP-45DP

Technical Specification	Electric Mode	Diesel Mode
Length over coupler	71 ft 6¼ in. (21,800 mm)	
Axle base	9 ft 2⅓ in. (2,800 mm)	
Wheel diameter	44 in. (1,118 mm), new 41⅞ in. (1,046 mm), fully worn	
Total weight	284,000 lb (128.8 tons)	
Axle load	71,000 lb (32.2 tons)	
Service speed	125 mph (201 km/h)	100 mph (160 km/h)
Converter type	IGBT, water cooled	
Nominal power at the wheels	5,360 hp (4,000 kW)	2,734 hp (2,040 kW), 8 cars
Train supply (HEP) capability	1,340 hp (1,000 kW)	
Starting tractive effort	71,000 lb (316 kN)	
Brake force (electric brake)	34,000 lb (150 kN)	
Brake resistor power	1,742 hp (1,300 kW)	

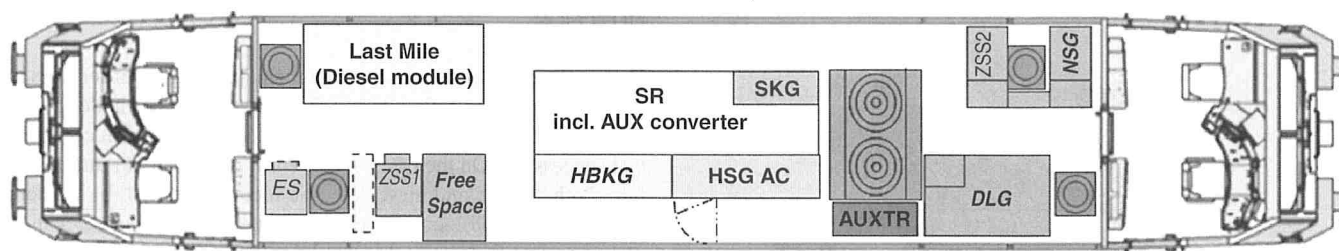


FIGURE 4 Equipment disposition in TRAXX AC3.

## DUAL-MODE LOCOMOTIVE DESIGNED FOR LAST MILE

Whereas the ALP-45DP was developed for the North American market, the requirements are quite different in Europe (3). There rail networks are mostly electrified, and important nonelectrified lines are increasingly fitted with catenary. Thus, diesel traction is often limited to less frequented secondary routes as well as sidings and terminals. In this situation, the design objective of a dual-mode locomotive in Europe is to maintain high performance for dense traffic on electrified main-line routes and to install only a small diesel engine for seamless services into sidings and terminals. Such a design concept has been implemented on the TRAXX AC3, a freight locomotive with 5,600 kW (7,500 hp) at the wheels in electric mode and 180 kW (240 hp) in diesel mode. The maximum tractive effort of 300 kN (67,400 lbf) is the same in both modes, allowing operation with the same train load also in diesel traction, although at much lower speeds, and without stopping for the switch-over. All electric propulsion and auxiliary equipment is placed into a central power pack, thus providing the necessary space for the last mile diesel module (see Figure 4). This module is designed as a self-contained, replaceable unit, complete with controls and cooling equipment. The TRAXX AC3 locomotive has, in addition, a boost function that provides supplementary power from batteries (see Figure 1). The diesel engine is a Deutz BR 2013 4V with a power rating at the shaft of 230 kW (310 hp). The engine fulfills the exhaust emission standard of Stage IIIB (similar to Tier 4), which has been required in Europe since January 2012.

## NEW DEVELOPMENTS IN DIESEL LOCOMOTIVES

It is possible to apply the concepts of multiengine propulsion, as used on the ALP-45DP, in diesel-electric locomotives. It was found that a four-engine solution with heavy-duty industrial engines is the most favorable for lowest initial and life-cycle costs. This finding led to the corresponding design of the TRAXX DE ME (BR 246) for the German railroad, which will use this locomotive for regional passenger services. The locomotive is shown in Figure 5. It has four Caterpillar C18 diesel engines compliant with the European Stage IIIB exhaust emission requirements. All engines feed into the DC link. It features an elaborate start and stop functionality so that engines can be powered selectively, according to the power requirements of the locomotive.

Such a solution can also fulfill the upcoming traction needs for new passenger services in North America on existing and future high-speed routes with speeds up to 125 mph. The basic requirements for such high-speed locomotives and coaches have been

specified by the Passenger Rail Investment and Improvement Act of 2008 Committee and are as follows:

- Service speed of 125 mph maximum,
- High traction power for fast train acceleration,
- Lowest possible weight to reduce fuel consumption,
- Low unsprung mass to reduce rail forces (particularly at high speed),
- Tier 4 exhaust emission standard (required starting in 2015),
- Low overall fuel and lube oil consumption, and
- Train supply (HEP) of minimum 600 kVA.

From the above list it is evident that a passenger diesel locomotive for high-speed services up to 125 mph has very different design requirements from those of a heavy freight locomotive. Lightweight designs with low axle loads, low unsprung mass, and high traction power are necessary to enable cost-efficient passenger operations. As a benchmark, axle loads in Europe for speeds up to 125 mph are limited to a maximum of 22.5 metric tons. A design example of such a truck is shown in Figure 6. Key design features are lightweight fabricated frame, primary and secondary coil springs, fully suspended drives including suspended brake discs, short axle base, flexible wheelset guidance vertically and laterally, and push-pull rod for transmitting traction and brake forces between the truck and car body.

Compared with a freight locomotive, a high-speed passenger locomotive also has a different operational profile and must provide redundant power for HEP in all modes of service, also at standstill.

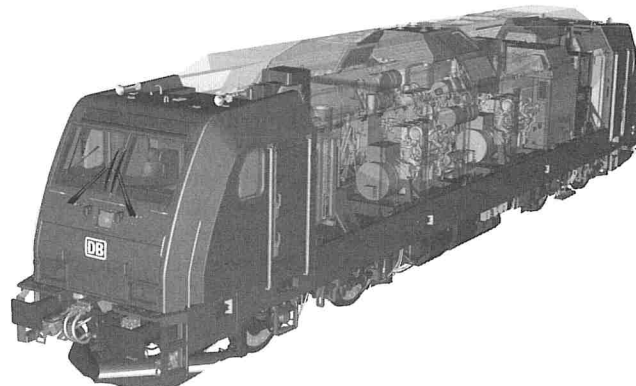


FIGURE 5 New TRAXX DE ME (multiengine) locomotive for German railroad DB Regio; it is powered by four heavy-duty C18 industrial engines.

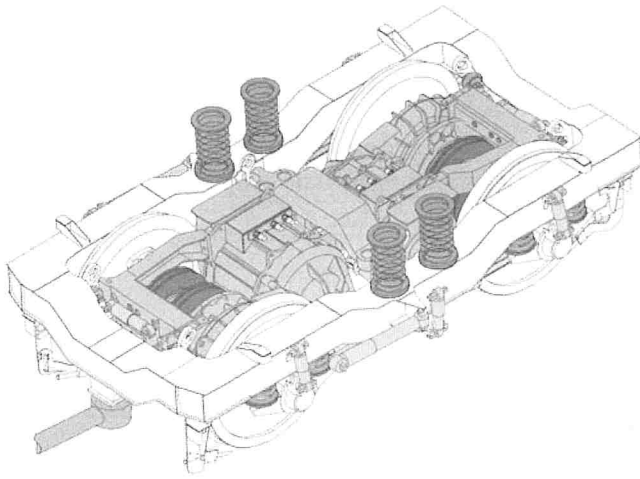


FIGURE 6 Design example of lightweight truck for 125-mph service.

In addition, frequent starts and stops must be accounted for. These requirements are well met with the multiengine propulsion concept. As with the above German locomotive, the advantages of multiengine propulsion compared with conventional diesel locomotives can be summarized as follows:

- Significantly lower overall locomotive weight compared with traditional single-engine solutions; with high-speed engines the weight savings are approximately 2 metric tons (on the basis of the four-axle TRAXX DE ME locomotive) for the Stage IIIB exhaust emission standard (the weight savings are of course considerably higher when compared with those of a single low- or medium-speed diesel engine);
- Faster acceleration of the engines and thus of the complete train;
- Proven industrial heavy-duty diesel engines that already today fulfill the exhaust emission standard of Tier 4 (the diesel engine assemblies are modular and can, for example, at future revisions, be replaced with new generation engines meeting future emission standards);
- Higher mission reliability with engine redundancy;
- Lower fuel and lube oil consumption with significant reduction of carbon dioxide emissions compared with single-engine concepts;
- Possibility to use engines of different suppliers without major modifications of the locomotive (therefore, second sourcing for the diesel engine is possible); and
- Substantially reduced maintenance and overhaul costs and increased spare parts availability during the product lifetime.

As all diesel engines feed into the same locomotive traction converter, the full tractive effort and HEP are always available, even if one or more engines are shut off. Engine start is by the alternator, eliminating the need for conventional starters and reducing stresses on batteries. The power conversion is similar to that of the ALP-45DP locomotive. Also here, a monocoque car body and fabricated trucks are prerequisites for lowering the overall weight and increasing the locomotive performance.

## OUTLOOK

During the next decades of commuter and passenger rail services in North America, the railroads and the industry will be challenged to enhance such future services and make them a long-term commercial success (4). On the basis of the North American rail infrastructure, both electric and diesel–electric locomotives will be needed. The dual-powered locomotive bridges the gap between electrified and nonelectrified networks. Applying the new technologies can contribute to productivity gains for the railroads so as to reach and maintain a competitive edge over alternative modes of transportation. Today, technologies for lightweight car bodies and trucks are available and proven in the United States and can also fulfill new standards, for example, crash requirements. Propulsion for traction is constantly being developed further, allowing new combinations of diesel, electric, and battery power sources, as seen in the new ALP-45DP and TRAXX AC3 developments. By taking these technologies forward, new diesel–electric passenger locomotives are also now feasible with multiengine designs, which fulfill Tier 4 requirements and have the potential of substantial fuel savings and exhaust emission reductions.

In future U.S. freight diesel locomotives, the above technologies would be much less targeted to reduce axle loads, but could be used rather to add a module for electric propulsion under catenary, for example, in tunnels and in regions sensitive to diesel exhaust emissions and engine-induced vibrations in track beds. In Europe the innovation is to add diesel propulsion to the existing electric locomotive fleets, whereas in North America the situation is reversed. Here the locomotives are basically diesel, and new technologies allow expanding the range of the locomotives with an additional electric propulsion package. In addition, the above technologies allow energy savings with power regeneration into batteries and ultracaps, which can be significant in stop-and-go traffic and shunting. Also, these technologies allow the combination of several diesel engines in a single freight locomotive with fuel-saving features as described above with the German TRAXX DE ME. Such savings can be considerable with medium-distance light freight trains hauled by a single locomotive, for example, in light intermodal applications.

## REFERENCES

1. Brugger, P., L. Schwendt, M. Spillmann, and J. Vitins. The Dual-Mode Locomotive ALP-45DP: An Innovation for the North American Market. *Railway Update*, Vol. 11–12, 2009, pp. 234–237.
2. Vitins, J. Reducing Energy Costs with Electric, Diesel and Dual-Powered Locomotives. *Proc., 2009 IEEE/ASME Joint Rail Conference*, Pueblo, Colo., March 4–5, 2009.
3. Vitins, J. The TRAXX Platform: A New Way to Build Electric and Diesel Locomotives. *Proc., 2008 IEEE/ASME Joint Rail Conference*, Wilmington, Del., April 22–23, 2008.
4. Vitins, J. High Speed Locomotive Development—A European Experience. *Proc., 2010 IEEE/ASME Joint Rail Conference*, Urbana, Ill., April 27–29, 2010.

*The Passenger Rail Equipment and Systems Integration Committee peer-reviewed this paper.*

# Development of the Next Generation of Intercity Corridor Bi-Level Equipment with Crash Energy Management

Eloy Martinez, Frances Nelson, Anand Prabhakaran, and Antony Jones

An innovative rail car procurement specification was developed for bi-level equipment. This specification was developed to fulfill the Passenger Rail Investment and Improvement Act of 2008 (PRIIA). PRIIA is a congressional mandate to the National Railroad Passenger Corporation, state departments of transportation, FRA, and passenger rail car builders and suppliers to develop the next generation of passenger rail equipment for intercity corridor service for speeds up to 125 mph. Technological innovations are to be used to improve safety incrementally, and components are to be standardized to the extent possible to leverage economies of scale to help revitalize the manufacturing base for domestic passenger equipment. This paper focuses on the technical discussions held by a cross section of key industry stakeholders to develop specification language for crash energy management features as an overlay on a fully compliant bi-level car design. The purpose of the paper is to widely disseminate the methodology and process used and to provide background information on the values chosen for individual car crush zone performance.

The Passenger Rail Investment and Improvement Act of 2008, Section 305 (PRIIA 305), is a U.S. congressional mandate that required the establishment of a Next Generation Corridor Equipment Pool Committee, composed of representatives of the National Railroad Passenger Corporation (Amtrak), FRA, host freight railroad companies, passenger railroad equipment manufacturers, interested states and, as appropriate, other passenger railroad operators. The executive committee was established in December 2009 and shortly thereafter created two subcommittees to develop procurement specifications that meet the intent of the U.S. congressional mandate and the means to purchase or lease the equipment. This paper provides some of the technical background associated with work conducted by the technical subcommittee and subgroups to develop a procurement specification for bi-level equipment that incorporates enhanced crash energy management (CEM) crashworthiness features.

The format of the paper follows the process used to develop the CEM section of the Next-Generation Corridor Equipment Specifi-

cation by the structures subgroup of the technical subcommittee. The most critical of the early decisions by the structures subgroup was to require CEM. After all industry stakeholders agreed that having CEM was appropriate, most of the effort was spent on determining how CEM should be implemented. Once the details were decided, it was necessary to confirm that the new requirements with CEM were indeed an improvement over existing requirements.

## BACKGROUND

The first meeting of the technical subcommittee was held on March 4, 2010. Subsequent deliberations led to the prioritization of vehicle types and the organization of technical working groups that included car builders, major subsystem and component manufacturers, and industry consultants.

The technical subcommittee decided that the first car type to be addressed would be a bi-level car capable of service in the western United States at speeds up to 125 mph. In addition to the bi-level car specification, a specification for a nonelectric locomotive for the same service environment would be addressed. It was also agreed that to complete the bi-level car specification in the time frame mandated by the executive committee it would be necessary to use the California Department of Transportation (Caltrans) C21 Corridor Car Technical Specification as a baseline document.

The first industrywide meeting of the technical working groups was held in Chicago, Illinois, on April 22, 2010. Vehicle prioritization (bi-level car and locomotive) and the desired completion date for the first car specification of July 2010 were communicated to the group. The July 2010 completion time equated to an extremely aggressive 3-month time frame for the specification development.

A series of groups were formed to address rail car design and the standardization mandate:

- Car system integration group,
- Structural group,
- Mechanical group,
- Electrical group,
- Interior configuration layout design group, and
- Vehicle track interaction-truck group.

To ensure that new equipment performance needs were captured, the groups included representation from all interested stakeholder groups. The objective of each group was to develop a list of equipment performance needs and define design, service, maintenance, and repair requirements to address those needs. Each group was asked to pool the collective knowledge from operators and designers of

---

E. Martinez, LTK Engineering Services, 10 Milk Street, Suite 701, Boston, MA 02108. F. Nelson, LTK Engineering Services, 100 West Butler Avenue, Ambler, PA 19002. A. Prabhakaran, Sharma & Associates, Inc., 5810 South Grant Street, Hinsdale, IL 60521. A. Jones, Voith Turbo Scharfenberg GmbH & Co. KG, Gottfried Linke Strasse 205, Salzgitter D-38239, Germany. Corresponding author: E. Martinez, emartinez@ltk.com.

---

*Transportation Research Record: Journal of the Transportation Research Board*, No. 2289, Transportation Research Board of the National Academies, Washington, D.C., 2012, pp. 47–55.  
DOI: 10.3141/2289-07



the equipment and components and to describe what worked well in service, what needed improvement, and what should be avoided in future designs.

A decision developed by the technical subcommittee was that the next generation of intercity corridor vehicles must be standardized to reduce the costs of manufacture and maintenance. Individual states may purchase the vehicles necessary to run service in the state, and Amtrak could be the operator and maintainer of the equipment. Standardization among procurements leverages economies of scale, potentially making the project more attractive to manufacturers. Performance standards are needed to allow equipment designed to the specification to operate under all conditions for any type of service with any type of consist.

In addition, it was desired that the specification use appropriate technological advances deemed to be truly next generation. CEM is an example of such a technological advancement from the perspective of structural crashworthiness.

It was the task of the structural group to develop specification language that included CEM. Overall, the work of the structural group focused on the following section of the base specification for modification: Chapter 2, References and Glossary; Chapter 4, Carbody; Chapter 6, Couplers and Draft Gear; Chapter 18, Materials and Workmanship; and Chapter 19, Test Requirements.

## CEM FOR NEXT-GENERATION CORRIDOR EQUIPMENT

The first task that the structural group addressed was whether in fact the Next-Generation Corridor Equipment Specification should include requirements for CEM. It was mandated that there be an improvement in technology, but would incorporating CEM requirements be feasible?

At the April 22, 2010, meeting in Chicago, FRA presented information to the technical subcommittee on changes occurring currently in the North American rail market. FRA cited, as an example, the recent

procurement by the Southern California Regional Rail Authority (SCRRA) Metrolink of new cab and coach car designs with CEM design features (Technical Specification for Southern California Regional Rail Authority, Contract No. EP142-06). FRA also pointed out the numerous waiver requests for equipment built to alternative performance standards for shared use that use this type of technology.

Figure 1 shows the comparison in performance of a conventional North American train and a prototype CEM-equipped train from one FRA full-scale impact testing program. CEM improves the performance in a head-on impact through controlled progressive collapse of energy absorbers while controlling the interactions of colliding and coupled equipment. The potential for either override or lateral buckling to occur is minimized through the controlled progressive collapse of structural features that make up the crush zone.

Application of CEM is standard for European transit, commuter, and intercity passenger rail car designs (Railway Applications—Crashworthiness Requirements for Railway Vehicle Bodies, EN 15227:2007). A quick survey of the car builders present confirmed that they all had some level of design experience implementing CEM in international markets. Therefore, the structural group agreed that the state of the industry was mature enough for inclusion of CEM because of recent advances in designs, fabrication techniques, and modeling capabilities.

## DEVELOPMENT OF CEM REQUIREMENTS

### Overall Concept

The next step was to determine how CEM should be implemented. The guiding principle that the group followed was to ensure that the introduction of the new equipment with CEM results in an improvement in safety in collisions and derailments. That is, under no circumstances should the introduction of CEM result in a decreased level of safety when compared with leaving the design based on conventional strength-based requirements.

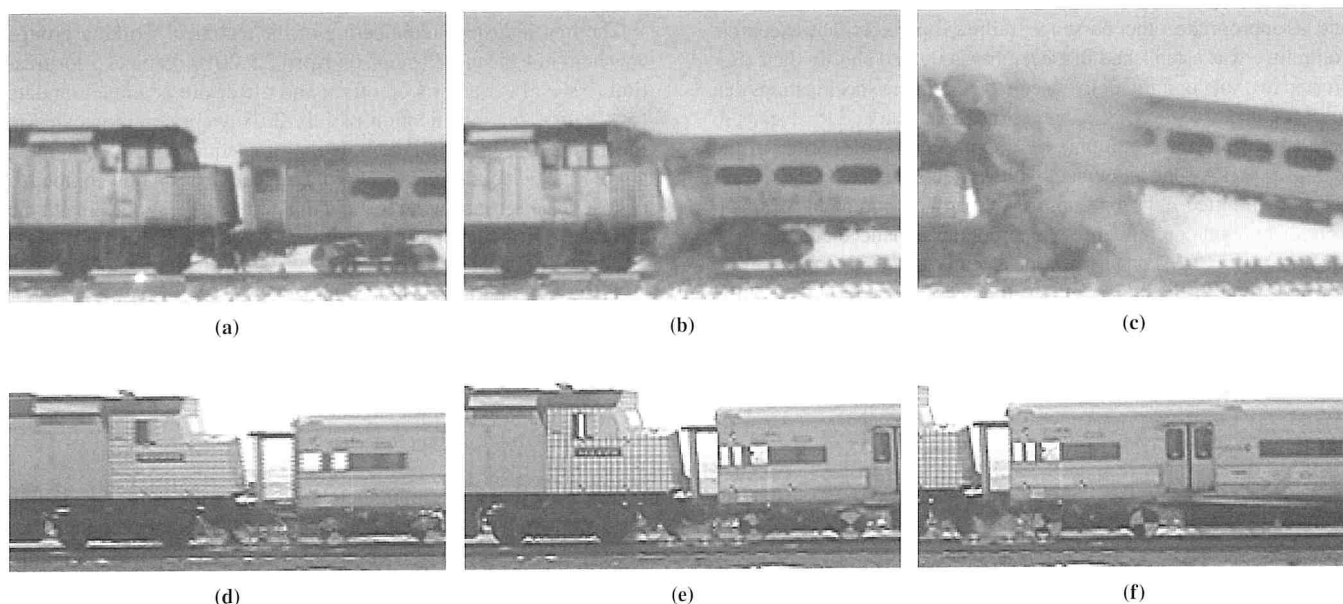


FIGURE 1 Train-to-train test comparisons between (a, b, and c) conventional equipment and (d, e, and f) CEM-equipped train (1).

Questions arose within the structural group, as well as within the technical subcommittee, as to the appropriate level of CEM that should be provided and whether reductions in the buff strength requirements could be achieved.

The requirement for compatibility and interoperability within an existing fleet considered for PRIIA 305 equipment is different from the discussions held by the Railroad Safety Advisory Committee (RSAC) Engineering Task Force (ETF), which was charged with developing a guideline for expedited waiver request of train set designs for mixed use on the North American rail network. In the RSAC ETF guidelines, options are allowed for alternative definition of buff strength and occupied volume strength. Such alternate buff force requirements are appropriate for train sets, for which the performance is defined at the train set level, and no intermixing of individual cars is expected.

However, vehicles developed through the PRIIA 305 bi-level specification are expected to be interoperable in mixed service with other and older Tier 1-compliant cars. Given the need to provide specification language for individual car types without previous knowledge as to how the equipment will be intermixed within existing equipment, the structural group decided that CEM would be provided as an overlay on a fully compliant car body structure (FRA, Department of Transportation, 49 CFR Part 238 et al., Passenger Equipment Safety Standards, final rule). This decision was presented to the full technical subcommittee for confirmation and was approved.

Subgroups within the structural group were formed to address different aspects of the design development and specification drafting. Each subgroup had a member from each interested industry stakeholder group. The three groups formed included a pushback coupler group, a car body structural group, and a group to address materials, workmanship, and testing requirements. Key questions and issues addressed by each group are discussed in the next sections in the order that information was generated.

The structural group agreed that a reasonable approach would be to define requirements at a car level in the manner that load enters the car. Therefore, discussions first focused on coupler requirements followed by the car body, an operator's survival space (this includes sloped nose), and finally other details in the specification that are affected by CEM.

The CEM approach proposed was to use pushback couplers in conjunction with car body-based energy absorbing elements to achieve the desired levels of protection.

## Couplers and Draft Gear

The first consideration when requirements for couplers were developed was defining typical service loads experienced to ensure that premature triggering would not occur. The specification had to define pushback coupler activation forces that are readily achieved by all vendors to ensure that couplers from various suppliers will actuate consistently in accidents, even when intermixed with each other, to distribute and control crush. In addition, because the equipment will spend some time comingled and intermixed with the existing fleet, the necessary stroke had to be defined to allow effective engagement with conventional cars.

Amtrak and Caltrans representatives agreed to provide common coupling and train makeup procedures for use as a starting point in developing appropriate activation force levels. In addition, typical

TABLE 1 Impact Scenarios Analyzed

Impact Simulation	Speed, km/h (mph)	Peak Coupler Force, kN (kips)
Locomotive into 1 car	6.4 (4)	1,810 (405)
	8.0 (5)	2,610 (584)
Locomotive into 5 cars	6.4 (4)	1,860 (417)
	8.0 (5)	2,620 (587)
Locomotive into 10 cars	6.4 (4)	1,820 (408)
	8.0 (5)	2,620 (587)
Locomotive + 5 cars into 5 cars	6.4 (4)	1,540 (345)
	8.0 (5)	2,120 (475)
Locomotive + 10 cars into 10 cars	6.4 (4)	1,540 (345)
	8.0 (5)	2,120 (475)

draft gear characteristics were provided by Caltrans and Bombardier Transport.

A number of simulations were performed to establish reasonable activation force levels (see Table 1). The impact conditions analyzed included a locomotive weighing 286,000 lb and a number of cab, coach, and service cars that weighed 150,000 lb. The moving train (whether a single locomotive or a locomotive-led train) struck a standing train with defined characteristics with the parking brakes applied. Friction was assumed to be 0.2 for the wheel-rail contact. The locomotive draft gear modeled was a Type 390 elastomeric draft gear. The passenger car draft gears modeled were WM6DP elastomeric draft gears. The draft gear characteristics are critical when coupling speed impacts in a one-dimensional (1-D) model are calculated.

Figure 2 depicts the results from the first impact condition defined in Table 1. Each curve represents the response of a car in the moving consist. The highest load experienced was on the lead coupler, which experienced a force of 1,810 kN (411 kips).

On the basis of the results presented, the structural group chose to set the activation force level for the pushback coupler to be 2,640 kN (600 kips), which is higher than the peak force level seen during a 5-mph impact. The energy absorption levels for the car body elements were defined such that the required energy to be consumed in a progressive controlled collapse of the car body structure would be the same whether for cab cars, coach cars, or service cars. This choice was to address the desire for developing standardized components such as energy absorbers that can be readily switched out among different equipment designs. This desire meant that the additional energy absorption nominally required for the cab car would have to be built in to the pushback coupler elements.

Subsequent to detailed deliberations about required capacity, practicality, and industry experience, the group has recommended an energy absorption capacity of 200,000 ft-lbf over 4 in. of stroke for pushback couplers in the coach ends, and a capacity of 700,000 ft-lbf over 14 in. of stroke for pushback couplers in the cab ends.

The structural group discussed the need for specifying a centering device for the couplers. Some of the centering issues discussed by the group follow:

- Would the weight of a Type H coupler require a substantial centering force?
- Would there be issues with coupling on curves?
- Should the mechanism be passive or active?
- Would the mechanism have to fit within the existing pocket geometry?



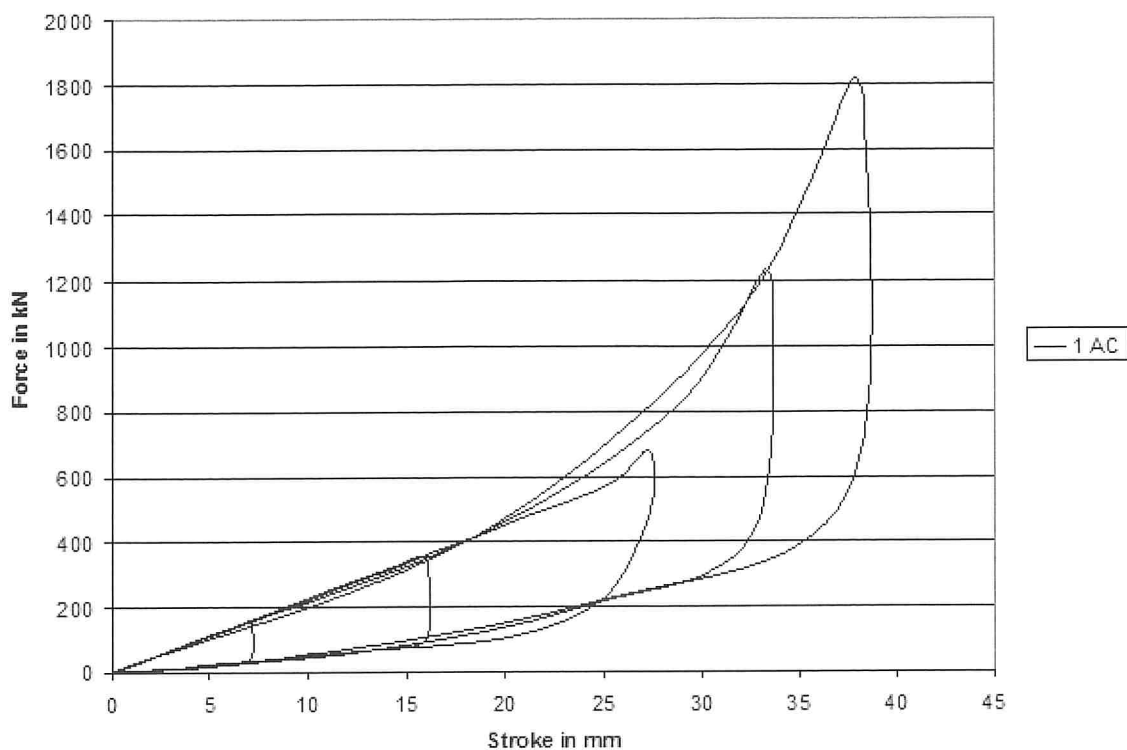


FIGURE 2 Force-stroke diagram of moving train set.

The group decided to include language for an active centering mechanism, but that it would not have to fit within the current coupler pocket design envelope. Language was added to the specification to clarify the needed stroke for pushback as well as using details of the SCRRA Metrolink specification to clarify that the pushback motion should not allow uncoupling of the equipment and the coupler carrier should not interfere with or affect the coupler pushback function.

The question of retrofit of existing equipment with pushback couplers remained open. Use of retrofitted pushback couplers can provide substantial improvements in car-to-car interactions during collisions. However, the group was unclear as to what the existing strength of vehicles is and whether the added potential pool of standardized components would outweigh the costs of repair after minor collisions, which would now affect every car end in the train as opposed to a single car end requiring more significant repair near the point of initial impact.

## Car Body

### Material

While the coupler group was evaluating trigger force levels, the car body subgroup discussed the potential for weight savings by allowing the use of alternative structural materials in the next-generation intercity corridor equipment. The question was posed to the technical subcommittee and, after discussions including the potential operators, was rejected. The potential operators were concerned with the performance of car shells constructed from aluminum in all environments that the corridor equipment would have to operate. Corrosion and fatigue were large concerns. Further, the cost for rou-

tine maintenance, inspection, and repair would potentially be greater because diversion from existing structures would result in the need for training and potentially new tools.

Once the car body material was decided, the car body subgroup discussed the following points:

- Should CEM performance be established on the basis of a crash scenario or by defining car level requirements for the CEM elements?
- How should the sequence of operation of the CEM elements be defined?
- What energy absorption levels are sufficient to ensure improved car level performance in minor accidents?
- What stroke, that is, the physical shortening of the car, should be allowed for the crush zone, and how many seats are lost as a result?
- Where should crush elements be located: in front of, under, or behind the engineer?
- What is the best means of defining a safe survival space for the engineer? Is a dedicated cab without walk-through capability and without train line doors necessary?

### Scenarios

Because of the uncertainty about placement of equipment constructed to the specification in the operating fleet, the structural group agreed that definition of a scenario would not be appropriate and instead the crashworthiness requirements would be defined at the car level.

The sequence of operation of the crush zone was developed assuming that the equipment coach and service cars would be placed within the consist and therefore squeezed between two planes. Cab cars were defined to interact with a rigid generic locomotive

geometry and function in a progressive fashion until the crush zone is exhausted.

### *Energy Values and Stroke*

After the coupler values were decided, a great deal of effort was given to determining the energy absorption levels and stroke of the car body crush zone.

A suggested starting point was to use the information in the SCRRA CEM specification. The concern with the specification was that it might result in force levels sufficiently high that excess weight would be required to strengthen the car structure to resist the loads.

As a consequence, the group agreed that it was preferable to specify both the stroke and the energy absorption level such that the average forces would potentially be elastic on a conventionally designed rail car.

The original value chosen by the structures subgroup for the car body crush zone stroke was 48 in. at each end. The value was eventually reduced to 24 in. of physical shortening of the car body structure at each end to minimize the loss of interior space for seats. This excludes the stroke of the pushback coupler, which occurs underneath the finished floor of the car.

The required energy absorption per car end was defined as 1,300,000 ft-lbf, independent of car type. This allows for standardization of individual crush elements of a modular design. In combination with energy absorption in the pushback coupler, this brings the total energy absorption of the coach cars to 1.5 million ft-lbf and the cab cars to 2.0 million ft-lbf.

The pushback coupler force is required to drop to zero before engagement of the car body energy absorbers to allow independent load path and reaction on the car body.

### *Occupied Volumes*

The next question addressed was how best to provide protection to the engineer. The idea of a dedicated cab was discussed. This included the concept of a shaped nose for the cab ends.

There are significant advantages to placing the crush elements outboard of the engineer, namely, reduction of acceleration and

therefore secondary impact forces in a collision. The concept of a shaped nose that could take advantage of the crush elements outboard of the engineer was discussed, but because of operational constraints was rejected for this equipment.

In typical service, a cab car may be placed in the middle of a consist, and at predetermined locations the train will be split in two. If the cab car is to be placed in the middle of a train, it is important to allow for passage from car to car through the cab and train line doors. This operational requirement was critical to one of the first potential users of the specification and therefore was agreed on by the structural group.

The placement of the crush elements was left to the discretion of the car builder, to be reviewed on initial proposal to the purchaser.

Language was agreed on for allowing the preservation of a defined safe survival space around the engineer, which on exhaustion of the crush zone will maintain a means for rapid egress from the cab. Figure 3 is a schematic of the required protected operator space.

### *Stress Levels*

It was decided to require that up to complete exhaustion of the crush zone, the stresses in structural members in occupied areas of the car—operator and passenger—remain elastic. It was agreed that local “hot spots” would be permissible provided the following three conditions are met:

- Plastic analysis of the model shows that the affected areas are small with plastic strain not exceeding 1%;
- With removal of the simulated load, there is no permanent set in the overall dimension of the occupant volume; and
- The function of the structure is not compromised.

### *Validation*

Further, it was agreed that validation of the analysis models would be accomplished through testing of crush and fuse elements. Once the submodels were validated, it would be necessary to assemble a complete three-dimensional (3-D) model of the car and perform

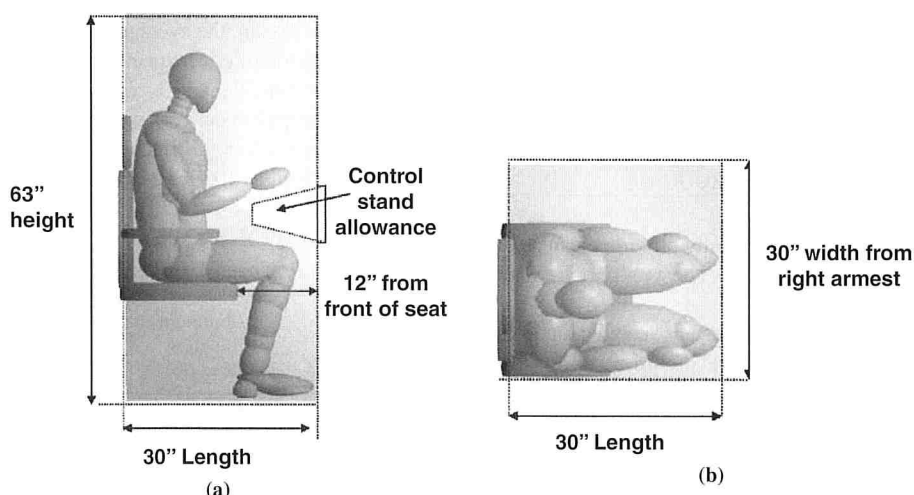


FIGURE 3 Operator seat clearance zone.

TABLE 2 CEM System Compliance Test Matrix

Test	Specification	Type	Car End	Level	Input Parameter	Criterion
Structural absorber energy absorption	4.21.4 and 4.21.5	Dynamic	All	Component	Energy absorption	Exceeds minimum required value
Coupler energy absorbed	6.9	Dynamic	All	Component	Energy absorption	Exceeds minimum required value
Coupler initiation load	6.9	Dynamic or quasistatic	All	Component	Initiation load	Exceeds minimum required value
Trigger, frangible element, fuse	4.21.4 and 4.21.5	Dynamic or quasistatic	All	Component	Load and failure mode	Within design range

explicit analysis of the car (by using a flat wall for coaches and service cars and impact into the generic rigid locomotive for cab cars) to demonstrate compliance with specification requirements.

Because of the maturity of the state-of-the-art in modeling, and following current practice, full-scale testing of a finished fabricated structure was not required.

### Additional Details

#### Testing

Extensive modifications were made to the car body testing section of the baseline C21 specification for purposes of this specification. This section of the paper will focus on the changes made to incorporate CEM into the design.

The structural group agreed to require dynamic or quasi-static testing as appropriate for each type of pushback coupler or structural energy absorber to validate the design performance. The car builder would furnish a purchaser with a CEM analysis and testing plan as part of the design review and approval process. Table 2 is the suggested CEM system test matrix.

For each element to be tested, the related part of the plan would have to include a description of the element to be tested, description of any required test fixtures, conditions under which the test will be conducted, and data to be measured.

Once the tests have been conducted, it was agreed that it would be necessary to evaluate the test results with respect to the pretest numerical predictions. Correlation of results should focus on the following:

- Peak force,  $\pm P\%$ ;
  - Average force,  $\pm A\%$ ;
  - Force and displacement versus time,  $\pm T\%$  at any particular time;
- and
- Modes of crush as predicted by analysis.

The group agreed that the car builder would provide meaningful and realistic correlation criteria for  $P\%$ ,  $A\%$ , and  $T\%$  for review and acceptance by the purchaser.

On satisfactory validation of the CEM element analytical models, the overall 3-D model of the car structure should be updated to reflect the test results, and compliance with the CEM section of this specification must be verified. If during the validation process it becomes apparent that some element of the CEM system would need to be redesigned, then redesigned components would be subjected to additional testing until satisfactory performance is demonstrated.

### Materials and Workmanship

The materials and workmanship section of the specification was updated to include reference to additional materials and fabrication techniques.

### Glossary and References

Chapter 2 of the C21 specification was updated to include information specific to use of CEM design features. Reference to the location for geometry of a rigid generic locomotive that was to be used for analysis of interaction with the front end of cab cars was defined.

### COMPARISON OF PROPOSED CEM REQUIREMENTS WITH EXISTING REQUIREMENTS

To assess the level of crashworthiness protection provided by a train set made up of equipment built to the specification, a number of 1-D collision dynamics calculations were conducted. Idealized force crush characteristics were developed and used as input for the simulations. Table 3 lists the proposed coupler characteristics used in the simulations. As described in the specification, the average slope of the force during pushback should be greater than or equal to zero.

The car body energy absorption characteristics were assumed to be similar for all cars with the minimum initiation load of 650,000 lbf and minimum energy absorption of 1,300,000 ft-lbf in a maximum 24-in. stroke. Again, the average slope had to be positive or zero to ensure activation of crush zones in a sequential fashion.

The impact conditions analyzed include a conventional locomotive-led train with passenger cars weighing 150,000 lbf and a locomotive weighing 260,000 lbf. The standing train comprises a CEM cab car followed by four CEM coaches and a conventional locomotive.

TABLE 3 Proposed Coupler Characteristics

Application	Minimum Initiation Load (lb)	Minimum Energy Absorption (ft-lb)	Minimum Push-Back Stroke (in.)
All, except cab end of cab car	600,000	200,000 @ 4 in.	9
Cab end of cab car	600,000	700,000 @ 14 in.	20

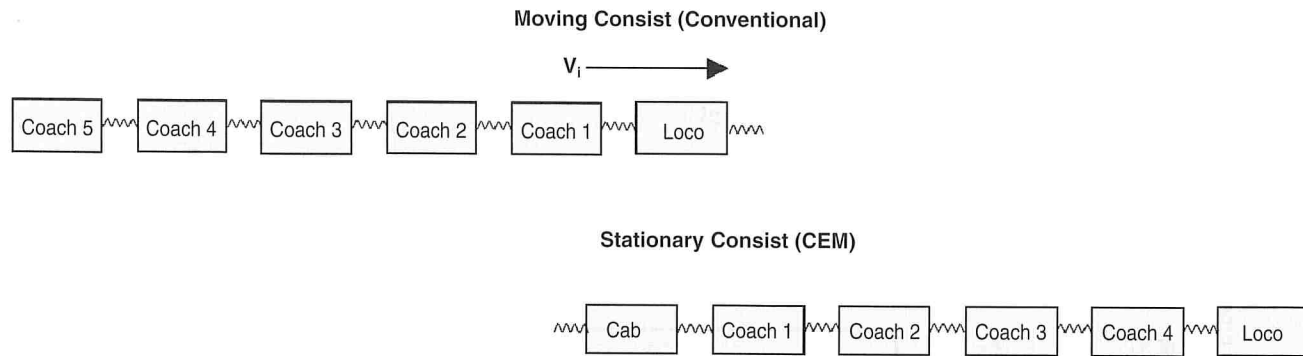


FIGURE 4 Schematic of 1-D collision scenario analyzed (loco = locomotive).

tive. The CEM cab and coach cars weigh 150,000 lbf, and the locomotive weighs 260,000 lbf. Three impact speeds were investigated: 18, 20, and 22 mph. The stationary consist has the brakes applied.

Figure 4 is a schematic of the 1-D collision dynamic model.

Figures 5 and 6 depict schematics of the idealized force crush characteristics used to describe the cab car and coach and service cars, respectively.

Results from the analysis on peak force experienced by the cab car are as follows:

- 2.2 million lbf at 22 mph,
- 1.9 million lbf at 20 mph, and
- 1.5 million lbf at 18 mph.

For each case analyzed, the CEM stroke was exhausted at the lead end of the cab car. It is suggested that the occupied volume strength of a fully compliant car loaded as defined through the structural energy absorbers on the car can easily resist dynamic loads between 2,000,000 and 2,500,000 lbf. Therefore, using the idealized force

crush characteristics under the collision conditions defined provide safe closing speeds between 20 to 22 mph. This provision is a significant improvement in safety compared with conventional equipment that has a safe collision speed under similar conditions of approximately 13 mph.

## SUMMARY

The technical specification for the next generation of intercity corridor bi-level equipment is the first specification that has been written and approved under PRIIA 305. This specification requires the car body structure to have CEM features overlaid on conventional Tier I design requirements. The structural group of the PRIIA Section 305 technical subcommittee was made up of members from Amtrak, FRA, state departments of transportation, car builders, major subsystem and component manufacturers, and industry consultants. The group was tasked with developing specification language for the inclusion of CEM for bi-level intercity cars capable of operational speeds of up to 125 mph with 5 in. of cant deficiency.

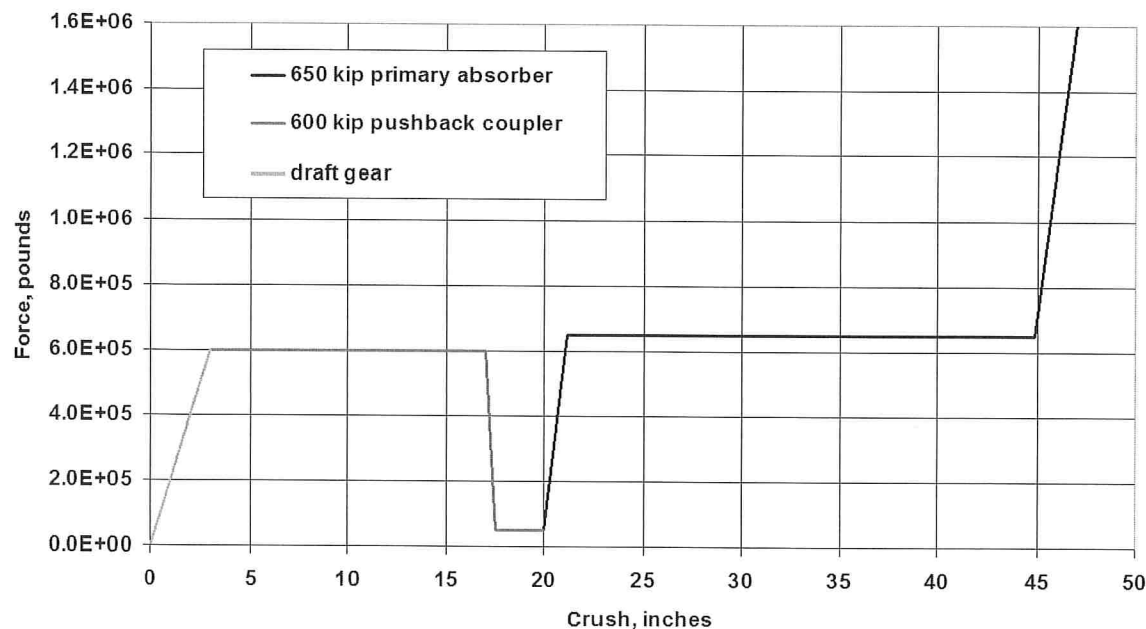


FIGURE 5 Idealized force crush characteristic of a cab car.

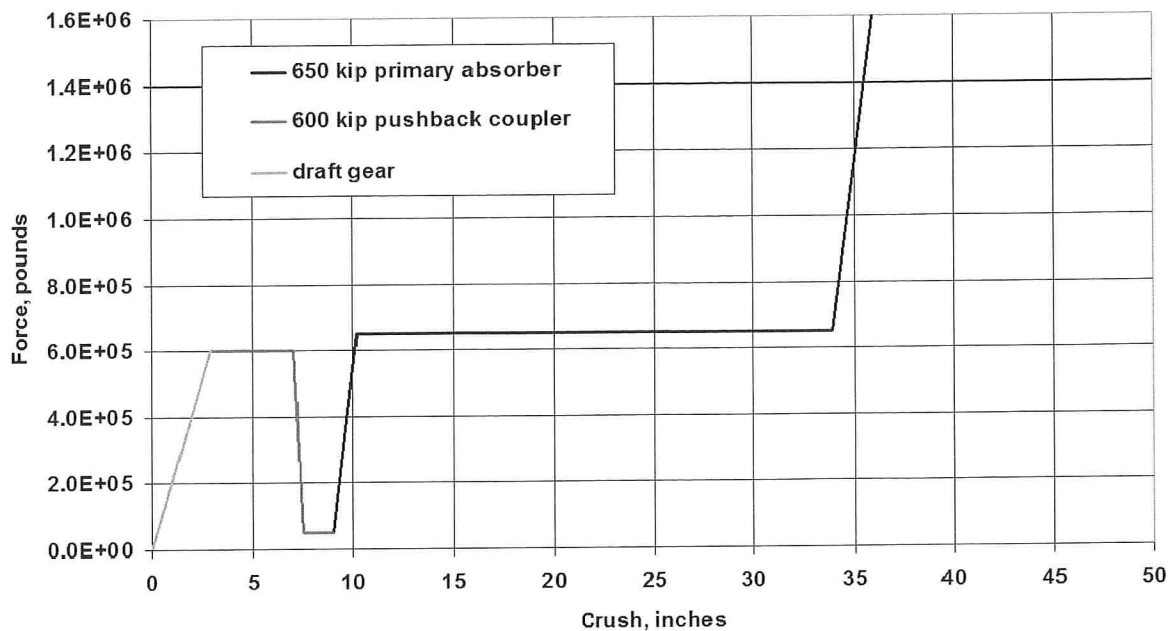


FIGURE 6 Idealized force crush characteristic of a coach or service car.

The goal of the structural group was to develop specification language that everyone could come to consensus on—the manufacturers needed to agree that they could bid on the procurement, the engineering consultants could agree that demonstration of compliance with the requirements was clear, and the form, function, and fit necessary for the end users was maintained.

The structural group met on weekly calls to discuss the research and analyses that individuals had completed during the week. It provided draft text for review by Amtrak on the sections of the specification affected by structural requirements.

The compatibility and interoperability requirements imposed by the PRIIA Section 305 Executive Committee required that the new equipment be fully compliant with all current federal regulations and industry standards. To introduce CEM, the crush zones were overlaid on top of a conventionally designed passenger rail car. The guiding principle was that the introduction of the new equipment with CEM result in an improvement in safety in collisions and derailments.

Several numerical simulations were conducted under severe coupling conditions to ensure that premature triggering of the pushback couplers would not occur. To ensure standardization from multiple parts suppliers, it was agreed that the trigger load must be defined consistently. The trigger load was established to ensure proper elastic performance of the draft gears up to coupling speeds of 5 mph. In the event that a pushback coupler was activated at speeds greater than 5 mph or in a grade crossing, requirements were developed to ensure that a damaged car or train set could be towed to a maintenance facility. Indicators are required such that a trainman walking the length of the train at a station can easily determine whether a pushback coupler has been activated. Care was taken to ensure ease of replacement of a damaged coupler. Maintenance costs should be no greater than use with conventional couplers.

The need for space at the ends of vehicles to introduce crush zone features required giving up some occupied interior floor space. How-

ever, careful attention to interior configuration layouts and placement of equipment closets, trash bins, bike racks, luggage racks, and similar uses that will not detrimentally affect the function of the crush zone can significantly reduce the potential loss of revenue seats. The current requirements are for implementation of crush zones at both ends of each car, with an allowance of up to 2 ft of physical shortening per end of the car.

For cab cars, the specification requires placing the cab on the second level of the equipment. There are requirements for a safe survival space for the operator in the event of an impact with a generic locomotive, and placing the cab on the second level will improve survivability.

The ability to walk through the train set is a critical requirement that affects the design of the cab car. Operators may wish to break a train in two at midpoints of a trip to increase the efficient use of equipment. In addition, development of the specification was based in part on the assumption that cab cars will be coupled in consist during operations. This results in the need to allow for passengers and crew to walk through the cab car when it is not in service as a cab car to reach the next car. The specification allows for this option to be retained. However, the specification also permits a shaped nose design in which the cab car is always placed at the end of a train set, such that walk-through capability is not required. There are advantages to such a design from the perspective of improved visibility from the cab and reduction of deceleration the operator may experience in a collision by placing the crush zone in front of the cab rather than under or behind the cab as is necessary with the flat end arrangement with walk-through capability.

The addition of CEM design to a Tier I-compliant car increases the safety and survivability of occupants under collision conditions in dedicated as well as mixed consist operations. The technology to implement this design strategy has matured during the past 10 to 15 years, and there is practical experience with it both domestically and internationally. The safety benefits offered by CEM designs

outweigh any potential costs or disadvantages. In addition, given the practicality of the solutions and the experience of the railroad industry with such designs, incorporating CEM features is the right way forward for future rail car acquisitions.

The specification developed as part of the PRIIA effort was approved by the PRIIA Technical Subcommittee and the PRIIA Executive Committee and is now the base specification for the next generation of bi-level cars to be procured in the United States.

## ACKNOWLEDGMENTS

The successful completion of this specification development, despite the large scope and the short timelines, was possible because of significant input from individuals at Amtrak, several state departments of transportation, FRA, car builders, major subsystem and component manufacturers, and industry consultants as part of the

PRIIA Section 305 Technical Subcommittee and members of the Passenger Rail Car Structural Subgroup. This work was primarily conducted pro bono by the participants to provide a refined procurement specification for intercity bi-level cars. The authors are extremely grateful to the many dozens of people who worked diligently on the several elements of this process to produce a practical and technically effective specification in a short time period.

## REFERENCE

1. *Justification for Inclusion of Crash Energy Management*. Submitted to the Technical Subcommittee of the Next Generation Corridor Equipment Pool Committee by the Structural Group, July 2010.

---

*The Passenger Rail Equipment and Systems Integration Committee peer-reviewed this paper.*



# Portable Emission Measurement System for Emissions of Passenger Rail Locomotives

H. Christopher Frey, Hyung-Wook Choi, and Kangwook Kim

The purpose of this study was to demonstrate a method for measuring passenger railroad locomotive emissions with the use of a portable emission measurement system (PEMS) based on rail yard load tests of three locomotives, including one GP40 and two F59PHIs. These locomotives have mechanically governed diesel prime mover engines (PMEs) with an approximately 3,000-hp output. Each locomotive has a head end power (HEP) engine that produces approximately 600 hp for generating electricity used in the passenger cars. The engine measurements were based on ultralow sulfur diesel fuel. Each engine was instrumented to measure manifold absolute pressure, engine revolutions per minute, intake air temperature, and exhaust concentrations of selected gases and particles. These data were used to quantify exhaust and fuel flow. The exhaust concentrations of nitric oxide, carbon monoxide (CO), carbon dioxide, hydrocarbons, and particulate matter were measured. The PMEs are operated at each of many throttle notch settings. For the HEP engines, three electrical loads were applied on the basis of power usage for one, two, and four passenger cars, respectively. More than 97% of the raw data survived a multistep quality assurance process. The data obtained from the PEMS for the main engines were found to be comparable on a fuel basis to data reported by others, particularly for oxides of nitrogen and CO. The key results from this work are the establishment of a simplified methodology for future tests and the development of baseline data.

There were 23,732 Class I locomotives in the United States in 2006 (1). The U.S. Environmental Protection Agency (EPA) estimates that locomotives consume approximately 4 billion gal of diesel fuel annually (2). Locomotive diesel engines are significant sources of air pollution (3–5); however, there are relatively few data on the emission rates of this source category. Furthermore, although freight locomotives typically have one large prime mover engine (PME) that is used to generate direct current for use in traction motors, many passenger locomotives have a separate head end power (HEP) engine that is used to generate 60 Hz alternating current to provide hotel services in the passenger train consist. There are few data on

HEP emissions. Freight locomotives include switchers and line haul. Switchers tend to be smaller, with PMEs of 2,000 hp or less, whereas line-haul freight PMEs can be approximately 4,000 to 6,000 hp each. There are some exceptions to this, such as recent introductions of switcher locomotives with multiple smaller engine “gensets” that can be turned on or off individually to match power demand. The passenger locomotive PME may typically be approximately 3,000 hp. Passenger locomotive HEP engines are approximately 600 hp but are not certified under the locomotive emission standard. Instead, the engines are subject to nonroad diesel engine rules.

The main source of PME emissions factor data is certification tests conducted to demonstrate that a particular engine make and model complies with an applicable locomotive standard. U.S. locomotive PMEs are subject to EPA emission standards promulgated in March 2008 (6). The standards apply to new and remanufactured engines. The standards contain a tiered approach to more stringent emission limits depending on the engine model year or date of remanufacturing. PMEs are operated at discrete throttle notch settings, including low idle, high idle for power takeoff, and eight notches that enable variation in useful output. Certification that a PME complies with the applicable standard is based on federal reference method (FRM) instrumentation and a test procedure in which the engine is run at steady state for each throttle notch position (7). The notch position steady state emission rates are weighted to represent a switcher or line-haul freight duty cycle, depending on the engine size and application. Such tests are expensive, are conducted under steady state conditions, and take place at a very limited number of facilities. There is significant cost in transporting the locomotives to the test facility as well as the lost revenue service during transport and testing and the cost of the test itself.

Owners and operators of locomotive fleets are under increasing scrutiny in regard to justification of operations on the basis of environmental considerations. To quantify the real-world emissions of locomotives, there is a need for a more flexible, convenient, and less costly approach to obtaining emissions data. These data can support evaluation of the effectiveness of engine remanufacture, alternative fuels, changes in operating practices, and comparisons of locomotives in a fleet. In turn, these data can support decisions on acquisition or remanufacturing of locomotives, selection of fuels, and improvement of operating procedures, for the purpose of reducing emissions. The quality objectives for these data are not as stringent as those for certification tests because the former may need only to provide insight on relative difference whereas the latter must enable comparison with allowable emission rates.

The use of a portable emission measurement system (PEMS) as a means to obtain data useful to a locomotive fleet owner is evaluated here. A PEMS can be easily installed for static measurements at a

---

Department of Civil, Construction, and Environmental Engineering, North Carolina State University, Campus Box 7908, Raleigh, NC 27695-7908. Current affiliation for H.-W. Choi: Greenhouse Gas Inventory and Research Center of Korea, 5th Floor, Gwanghwamun Office, 163 Sinmunno 1-Ga, Jongno-Gu, Seoul 110-999, South Korea. Current affiliation for K. Kim: Clean Fuels and Technologies Division, Department of Sanitation of New York, New York, NY 11377. Corresponding author: H. C. Frey, frey@ncsu.edu.

*Transportation Research Record: Journal of the Transportation Research Board*, No. 2289, Transportation Research Board of the National Academies, Washington, D.C., 2012, pp. 56–63.  
DOI: 10.3141/2289-08

rail yard and can also be installed on board a locomotive for over-the-rail measurements (planned for a subsequent study). Although PEMSs have been applied extensively for measurement of in-use emissions of cars, trucks, and nonroad vehicles such as construction equipment, they have not yet been widely applied to locomotives.

The objectives of this paper are to (a) develop and apply a methodology for assessing the activity, fuel use, and emission rates for locomotive PME and HEP engines; (b) measure emission levels of locomotive engines with a PEMS; and (c) evaluate the use of a PEMS as an alternative to engine dynamometer measurement and identify its comparative strengths and limitations.

## TECHNICAL APPROACH

The technical approach includes (a) study design, (b) PEMS instrumentation, (c) field data collection, (d) quality assurance and quality control, (e) data analysis, and (f) benchmark comparisons.

### Study Design

Field study design includes specifying which engines are to be tested, when they are to be tested, what fuel will be used, what type of duty cycle will be performed, and who will operate the locomotives. The study design depends on the study objectives. In this case, the objectives are to obtain a baseline characterization of PME and HEP emissions for three locomotives used in passenger rail service between Raleigh and Charlotte, North Carolina. The data are needed to assess whether and to what extent emissions differ when the three locomotives are being compared and to identify priorities (if any) among the three locomotives for future emission reduction efforts.

The selected locomotives are a GP40, NC1792, and two F59PHIs, NC1755 and NC1797, owned by the North Carolina Department of Transportation (DOT). Each locomotive has a PME used to provide direct current electric power for propulsion, and a HEP engine used to generate alternating current power for "hotel services" in the passenger train (8). Emissions for each engine from each locomotive were measured, for a total of six engines.

The specifications of the PMEs and HEP engines of the three locomotives are summarized in Table 1. Measurements are based

on ultralow sulfur diesel fuel. The three PMEs are two-stroke engines. Two-stroke engines are known to burn more lubricating oil than four-stroke engines, which may be a factor that influences emission rates.

## Portable Emission Measurement System

The PEMS used is the OEM-2100 Montana system manufactured by Clean Air Technologies International, Inc. (9). The Montana system comprises two parallel five-gas analyzers, a particulate matter (PM) measurement system, an engine sensor array, and an onboard computer.

The pollutants, which include oxygen, hydrocarbons (HC), carbon monoxide (CO), carbon dioxide (CO<sub>2</sub>), nitric oxide (NO), and PM, are measured by using the following detection methods:

- Nondispersive infrared (NDIR) to measure HC, CO, and CO<sub>2</sub> (10, 11);
- Electrochemical cell to measure NO (for diesel engines, NO<sub>x</sub> is composed of approximately 92% NO by volume) (12, 13); and
- Laser light scattering to measure PM, with measurement ranging from ambient levels to low-double-digits opacity (10, 11).

The performance of the Montana system has been verified in comparison with a laboratory-grade chassis dynamometer measurement system (14–17). The coefficients of determination ( $R^2$ ) exceeded .86 for all pollutants, thus indicating good precision. The slopes of parity plots for CO, CO<sub>2</sub>, and NO ranged from 0.92 to 1.05, indicating good accuracy, and ranged from 0.62 to 0.79 for HC. The bias for HC is a well-known result of the NDIR detection method (18). The PEMS is calibrated in the laboratory by using a cylinder gas and in the field is periodically recalibrated to ambient air to prevent instrument drift. The PEMS used here has been used for measurements of a wide variety of vehicles, including cars, trucks, and construction equipment and for a variety of fuels, including gasoline, E85 ethanol, ultralow sulfur diesel fuel, and biodiesel. Thus, the PEMS is applicable to a wide variety of vehicle and fuel types (19, 20).

Intake airflow, exhaust flow, and mass emissions are estimated with a method reported by Vojtisek-Lom and Cobb (11). The data needed for these estimates include manifold absolute pressure (MAP), engine revolutions per minute (rpm), and intake air temperature (IAT). A temporarily mounted sensor array is used to measure these three parameters.

## Field Data Collection Procedure

The measurements reported here were conducted under static conditions at a local rail yard in Raleigh, North Carolina. PEMS installation involves the following connections: (a) installing MAP, rpm, and IAT sensors; (b) connecting exhaust gas sample lines from the exhaust duct to the PEMS; and (c) providing power to the PEMS. Installation is facilitated by identifying in advance the details of these connections and providing sufficient time for rail yard mechanics to fabricate a MAP port and an exhaust sample line port for the duct of the PME and to identify a source of power. For rail yard tests, shorepower can be used. The installation process on the day of the tests takes about 2 h for the PME. The MAP sensor is connected to a fabricated port on the airbox of the engine. The rpm sensor is based on an optical device that detects the reflection of

**TABLE 1** Specifications of Prime Mover and HEP Engines of Tested Locomotives

Item	NC1792 (GP40)		NC1755 and NC1797 (F59PHI)	
	Prime Mover	Head-End Power	Prime Mover	Head-End Power
Engine make	EMD	Cummins	EMD	CAT
Engine model	16-645E3	KTA19	12-710G3	3412
Number of strokes	2	4	2	4
Number of cylinders	16	6	12	12
Displacement (L)	169	19	140	27
Horsepower (hp)	3,160	600	3,200	625

NOTE: NC1755 and NC1797 are model F59PHI locomotives built in 1998 and 1997, respectively. NC1792 is a model GP40 locomotive built in 1968. The prime mover engine and head end power 2 engines of the GP40 were rebuilt in 1992 and 2005, respectively.

light from reflective tape that is placed on a pulley wheel that rotates at the same rpm rate as the engine. IAT is measured with a thermocouple. The key installation steps include removing an airbox port and replacing it with an identical one that has a barb fitting for the MAP sensor, locating an appropriate position for the rpm sensor and reflective tape, finding a location in the air intake path for the IAT sensor, connecting the exhaust duct port and the exhaust tubing, routing wires and tubes between the locomotive and the PEMS unit, and connecting the PEMS to shorepower. For the HEP engine, the procedure is similar. However, instead of an exhaust duct, there is a tailpipe similar to that found on a truck for these three locomotives, and thus a standard exhaust sample probe used routinely with trucks is easily installed in the tailpipe.

After installation, the PEMS and engine were warmed up for 45 min. After the warm-up, the PME was run at Notch Position 8 for a period of approximately 3 min, after which the engine was returned to idling. During testing under load, the electrical power produced by the DC generator connected to the PME was dissipated in an electrical resistance grid that is referred to as the dynamic brake grid. There are cooling fans above the grid that are used for forced-air cooling. However, the grid is not intended for sustained operation at high electrical current. To prevent overheating, operation at Notches 6 through 8 was limited to 3 min. The load test at each of these notches was immediately followed by a period of idling to allow the grid to cool for 5 min. Thereafter, testing occurred sequentially for Notch Positions 5, 4, 3, 2, 1, and idle, without any intermediate idling.

The HEP was run at multiple electrical loads for a period of approximately 10 min per load. The electrical load conditions were none, low, medium, and high. The loads were imposed by attaching passenger rail cars and operating the lighting and air conditioning in each. Thus, the low, medium, and high loads correspond to the combined space conditioning and lighting loads for one, two, and four passenger cars, respectively. Voltages and currents were measured to estimate the electrical loads. During data collection, exhaust gas concentrations and engine data were recorded on a second-by-second basis.

The rail yard testing is the first step in a series of longer-term research tasks. Later, over-the-rail tests with PEMS are planned, to enable comparison of results under rail yard and over-the-rail conditions.

### Quality Assurance and Quality Control

The measured data are screened to check for errors. If errors are identified, they are either corrected or the data set is not used for data analysis. Details of the quality assurance procedures are given by Frey et al. (20). Three of the most common types of errors or problems are briefly described.

On occasion, an invalid reading is obtained for engine rpm from the optical sensor. The two-stroke PMEs typically operate between 250 and 950 rpm. Values outside this range are considered to be invalid and are removed before further data analysis.

"Freezing" refers to situations in which a value that is expected to change dynamically on a second-by-second basis remains constant over an implausibly long time period. On occasion, the gas analyzer output fails to update and appears to be frozen at a constant value.

Each gas analyzer is referred to as a "bench." Most of the time, both benches are in use. Each gas analyzer bench is "zeroed" on a staggered schedule every 15 min. While being zeroed, the gas analyzer will intake ambient air instead of tailpipe emissions. There-

fore, most of the time, the concentration measurements from each of the two benches can be compared. When the relative error in the concentration measurement between both benches is within a predetermined maximum allowable discrepancy (MAD), and if no other errors are detected, then an average value is calculated on the basis of both of the benches. However, if the relative error exceeds the MAD because of a problem, such as a leak in the sample line, overheating, or sampling pump failure, then only data obtained from the other bench are used.

### Data Analysis

Measured data are analyzed to estimate average mass per time fuel use and emission rates for each throttle notch position. Emission rates for each throttle notch position were also estimated on the basis of mass per gallon of fuel consumed. Weighted average emission factors were estimated for the line-haul freight locomotive cycle. The analysis method used here is similar to methods developed for other nonroad vehicles, such as construction equipment (21).

## RESULTS

Results include the field data collection schedule, quality assurance, comparison between engines, and comparison with independent data.

Field data collection took place during 2008. The PMEs were tested for locomotives NC1755 and NC1792 in March. In July, the PME of locomotive NC1797 was tested. Measurements were made in July on the HEP engines of all three locomotives. The time taken to install the PEMS, including the MAP, rpm, and IAT sensors; the exhaust sample line; and connections to shorepower, was about 2 h for the PME. The PME measurements took place during a period of about 1 h. After the PME tests, the sensors and exhaust lines were relocated to the HEP; the relocation took about an hour to do. The HEP measurements were done during a period of about an hour. Thus, tests on both the PME and HEP were completed in 1 day. Data analysis and reporting typically took about 5 days for each day of field measurements.

### Quality Assurance

On average, 97% of the raw second-by-second data were valid. Unusual engine rpm (which occurred only for one engine), gas analyzer freezing, and interanalyzer discrepancy accounted for, on average, the loss of 0.8%, 0.7%, and 1.2% of the raw data, respectively.

### Prime Mover Engines

The engine rpm for each of the three PMEs was approximately the same for a given notch position. At idle, the engines operate at approximately 250 rpm; at Notch 8 they operate at approximately 900 rpm. MAP varies from approximately 103 to 276 kPa, depending on the notch positions and the engine.

The exhaust concentrations for NO and CO<sub>2</sub> tend to increase as the notch position increases. For example, as shown in Figure 1, the NO concentration for the GP40's EMD16-645 engine increases monotonically from 164 to 1,555 parts per million (ppm) between idle and Notch 8. McKenna et al. reported similar NO emission

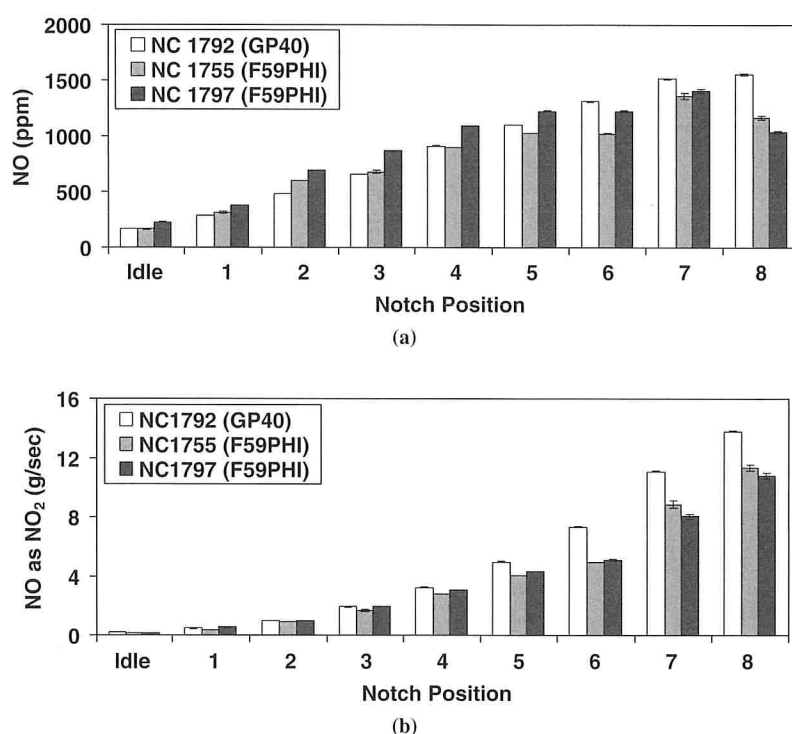


FIGURE 1 (a) Average NO concentration and (b) emission rate for prime mover engines of GP40 and F59PHIs versus throttle notch position.

concentrations for each notch position for an EMD16-645 engine (22). The NO and CO<sub>2</sub> concentrations for the EMD12-710 engines of the F59PHIs increase from idle to Notch 7 and are lower at Notch 8 than at Notch 7. NO<sub>x</sub> formation depends strongly on a combustion temperature (23). The NO concentration of NC1755, which has an EMD12-710 engine, increases from 160 to 1,360 ppm between idle and Notch 7. At Notch 8, the NO concentration is 1,170 ppm. The observed concentrations for HC and CO are less sensitive to notch positions in part because the concentrations are often at or below the detection limit of gas analyzers. The observed levels were between 0.6 and 19 ppm for HC and 0.0008 and 0.085 volume percent for CO. However, the detection limit is sensitive enough to make inferences as to whether the emission rates are at or above Tier 0+ or Tier 1+ levels, which would be required to be achieved as a result of a later engine rebuild. Diesel engines typically have very low emissions of CO and HC because they operate with high air-to-fuel ratios. CO and HC mass emission rates are not usually a main point of concern when diesel engines are evaluated; rather, the focus is on NO<sub>x</sub> and PM.

The NO emission rates increase monotonically between idle and Notch 8. For example, the average NO emission rates for NC1755 range from 0.13 to 11 g/s between idle and Notch 8. The NO emission rates for the older EMD16-645 engine are generally higher than for the two EMD12-710 engines; for example, the NO emission rate at Notch 8 is 22% to 28% higher than for the EMD12-710 engines.

Fuel-based emission factors, given in Table 2, were calculated on the basis of a carbon balance of the exhaust components, molar ratio of exhaust components to CO<sub>2</sub>, and fuel carbon content. Cycle average emission factors were estimated on the basis of the line-haul freight locomotive duty cycle. The EPA duty cycle includes a mode for dynamic braking that was not part of the stationary load test procedure. The percentage of time assigned to dynamic braking

in the EPA cycle was assigned to the idle mode. The cycle average emission rate is based on total emissions divided by total fuel use for the cycle.

The NO<sub>x</sub> emission rate for the NC1755 F59PHI locomotive varies from approximately 170 to 240 g/gal among the notch positions. Although a large percentage of time for the freight line-haul cycle occurs in the idle mode, most of the fuel consumed (65% to 67%, depending on the engine) is estimated to take place at Notch 8. The mass per time rate of fuel consumption at Notch Position 8 is approximately two orders of magnitude higher than at idle mode. Given the high fuel consumption rate at high engine load, the fuel-based cycle average emission factors are most influenced by the modal emission factors at high engine load. The cycle weighted average NO emission rate is 220 g/gal.

The two F59PHI locomotives have similar cycle average NO<sub>x</sub> and opacity-based PM emission rates. Apparent differences in cycle average HC and CO emission rates are not significant. The average concentrations are below the detection limit of 13 ppm for HC and 0.012 volume percent for CO. The GP40 has a similar average opacity-based PM emission rate compared with the F59PHIs, and a somewhat higher average NO emission rate. The average HC and CO emission rates for the GP40 are within the range of those for the F59PHIs.

A comparison of the fuel-based emission factors from the three locomotives versus the range and average of values reported by EPA is given in Table 2 (21, 24). For NO<sub>x</sub>, the measured emission factors are comparable with the range reported by EPA, although they tend to be at the low end of the range. The PEMS measures NO but not total NO<sub>x</sub>. Total NO<sub>x</sub> is typically about 92% NO. Thus, the measured emission factor is increased by a ratio of 1.087 (1/0.92), leading to values of 230 to 260 g/gal. These values are within the range of the EPA data. Overall, the NO emission measurements are deemed to be reasonable.

TABLE 2 Fuel-Based Emission Factors Based on Notch Position for GP40 and F59PHI Locomotive Prime Mover Engines

Locomotive Number, Model, and Engine	Notch Position	Fuel Use (%) <sup>a</sup>	NO as NO <sub>2</sub> (g/gal)	HC <sup>b</sup> (g/gal)	CO <sup>b</sup> (g/gal)	Opacity-Based PM (g/gal)
NC1792, GP40, EMD16-645	Idle	2.8	240	14	52	3.5
	1	0.8	260	9.3	34	2.0
	2	1.9	230	8.2	16	2.5
	3	2.8	240	5.9	8.4	2.0
	4	3.7	260	3.1	3.9	1.8
	5	5.0	260	1.6	2.5	1.6
	6	7.4	260	0.6	5.1	1.3
	7	8.8	260	1.6	6.4	1.1
	8	67	230	3.5	14	1.5
Cycle average for NC1792 <sup>c</sup>	—	—	240	3.5	13	1.6
NC1755, F59PHI, EMD12-710	Idle	2.1	240	1.7	13	5.3
	1	1.0	170	1.2	5.9	2.4
	2	2.3	200	5.8	3.3	2.6
	3	3.5	190	4.4	1.1	2.1
	4	4.6	200	1.6	5.3	2.3
	5	5.7	210	1.3	5.9	2.2
	6	7.6	200	0.4	7.4	2.1
	7	8.8	230	2.7	19	1.5
	8	65	220	1.8	12	1.2
Cycle average for NC1755 <sup>c</sup>	—	—	220	1.9	11	1.6
NC1797, F59PHI, EMD12-710	Idle	2.5	250	14	5.5	9.4
	1	1.4	190	7.8	13	4.2
	2	2.2	210	7.0	11	2.6
	3	3.4	230	6.9	2.2	2.2
	4	4.5	230	4.1	5.3	1.7
	5	5.4	230	1.2	12	1.5
	6	7.1	210	1.2	14	1.4
	7	8.9	200	2.8	21	1.1
	8	65	200	5.1	16	1.1
Cycle average for NC1797 <sup>c</sup>	—	—	210	4.7	15	1.5
EPA minimum <sup>d</sup>	—	—	220	3.1	11	5.0
EPA maximum <sup>d</sup>	—	—	320	15	51	8.5
EPA fleet average <sup>d</sup>	—	—	260	10	32	6.3

NOTE: — = not applicable.

<sup>a</sup>The fraction of fuel use for freight line-haul cycle is calculated on the basis of time-based fuel use rate and line-haul duty cycle. The fraction of fuel use is adjusted in that dynamic braking is assigned to idle mode.

<sup>b</sup>Italic numbers indicate emission rates based on exhaust concentrations that are below the detection limit of gas analyzers. The detection limits for HC and CO are 13 ppm and 0.012 volume percent, respectively.

<sup>c</sup>This is a cycle average emission factor based on adjusted line-haul cycle. Emission factors based on passenger cycle are similar to those for line-haul cycle within 5% difference.

<sup>d</sup>Data based on EPA reports converted to a fuel basis (7, 24).

For HC, the measured emission factors are low when compared with the weighted average from the EPA reported data. For one F59PHI, the measured average emission factor is less than the minimum value estimated on the basis of EPA's data. However, the measured average emission factors for the GP40 and the other F59PHI are slightly higher than the minimum value. The HC measurement is based on NDIR, which is known to be accurate for straight chain hydrocarbons but is less accurate for more complex molecules (such as aromatics). Typically, NDIR HC measurements may need to be adjusted with a bias correction of two or more to correspond to the actual total hydrocarbon load in the exhaust (18). If a factor of two adjustment is applied here, then the measured emission factors would be 3.8 to 9.4 g/gal. Although still at the low end of the range of data inferred from EPA's report, these values are consistent with the benchmark data.

As noted earlier, the CO exhaust gas concentrations are typically below the detection limit of the gas analyzers and thus are subject to uncertainty. Nonetheless, the average emission factors estimated from the measurements are comparable with the data reported by EPA.

The opacity-based PM measurements are clearly lower than the benchmark data. As noted earlier, these measurements are based on a light scattering laser photometer detection method. These measurements are useful for relative comparisons of data obtained by using the same method but are not appropriate for characterization of the absolute total emissions. The data here suggest that the three locomotives have comparable PM emission rates. The data showed that the opacity-based PM emission rates are approximately a factor of four lower than the average estimated fuel-based rate based on data reported by EPA.

The brake-specific fuel consumption (BSFC) for these locomotives is not known because there is no measurement of shaft torque nor is there an electronic control module that reports estimates of such data based on engine maps. Thus, it is not possible to estimate emission factors in units of grams per brake-horsepower-hour (g/bhp-h) directly. However, it is possible to make an estimate of g/bhp-h emission factors for Notch 8 by using a typical value for BSFC. EPA reports a typical BSFC of 0.048 gal/bhp-h (21, 24). Measured BSFC at Notch 8 is



**TABLE 3 Average Brake-Specific Emission Factors for Throttle Notch 8**

Locomotive Number, Model, and Engine	NO (g/bhp-h)	HC (g/bhp-h)	CO (g/bhp-h)	Opacity-Based PM (g/bhp-h)
NC1792, GP40, EMD16-645	11	0.17	0.65	0.073
NC1755, F59PHI, EMD12-710	11	0.09	0.58	0.060
NC1797, F59PHI, EMD12-710	8.8	0.25	0.76	0.054

NOTE: Brake-specific emission factors were estimated on the basis of fuel-based emission factors in Table 2 and BSFC of 0.048 gal/bhp-h (7, 22, 24, 25).

reported as 0.048 gal/bhp-h for EMD16-645 and EMD12-710 engines (22, 25). On the basis of this value of BSFC, estimated brake-specific emission factors for the Notch 8 position are shown in Table 3.

EPA reported 14 and 11 g/bhp-h of line-haul duty cycle average NO<sub>x</sub> emission rates for EMD16-645 and EMD12-710 engines, respectively (21). On the basis of other data, NO<sub>x</sub> emission rates for EMD16-645 have been reported in a range from 12 to 13 g/bhp-h (26). The estimated Notch 8 emission rate for the EMD16-645 is slightly less than these values and varies for the two measured EMD12-710 engines between 9 and 11 g/bhp-h. The older engines are not required to comply with the new standards until such time as they undergo engine rebuilds. Furthermore, the measurements conducted here are intended to evaluate relative differences between notches and engines and are not an FRM.

For HC, NC1792 and NC1797 have emission rates that are comparable with the Tier 3 standard of 0.3 g/bhp-h, by taking into account the known bias in NDIR measurements of total HC. For CO, it is likely that all three locomotives can comply with any of the tiers of the locomotive standards. For PM, there is considerable uncertainty given the semiquantitative nature of the measurement method. There is not yet a standardized method for measuring PM with portable instruments. This area is one of active and ongoing research among PEMS developers.

Rebuilds of these engines, which were planned as of the date that these measurements were made, would need to focus on significantly

reducing NO<sub>x</sub> emissions to achieve the requirement for Tier 0+ compliance that would be triggered by a rebuild.

## Head End Power Engines

Table 4 shows the fuel rate and fuel-based emission rates for the HEP engines versus electrical load. The variation in electrical load for a given number of passenger cars from one test to another is due to variability in ambient temperature and solar irradiation, which affects the cooling load. At a load of "none," some power was consumed to maintain the battery charge of the locomotive's batteries.

The rate of fuel use, CO<sub>2</sub> emissions, and NO emissions increases as electrical load increases. HC emission rates for diesel engines typically depend less on load and more on air-to-fuel ratio (27). The CO emission rate tends to be highest at no load, which is a relatively inefficient operating condition. Fuel-based PM emission rates are approximately similar among various loads. Although the Cummins KTA19 engine, which was rebuilt in 2005, has higher electrical loads than the CAT 3412 engines, the fuel use and emission rates were lower.

As expected, the fuel use and CO<sub>2</sub> emission rates of the two CAT 3412 engines are similar for comparable loads. The NO emission rates are of similar magnitude but appear to be slightly higher for NC1755 than for NC1797. The HC, CO, and PM emission rates are comparable in magnitude.

It has not been possible to identify published data on the same make and model of HEP engines. As a benchmark, EPA certification data for similar size engines manufactured by Cummins and CAT in 2003 were compared with fuel-based emission factors here. Nonroad engine emission data reported by EPA for 2003 were the most recently available. Because EPA data were reported in grams per brake-horsepower-hour, grams-per-gallon emission factors were calculated on the basis of engine-specific BSFC.

The Cummins 3CEXL019 is a 19-L nonroad diesel engine with BSFC of 0.0427 gal/bhp-h. The certification emission rates are 140, 9, 36, and 5.2 g/gal for NO<sub>x</sub>, HC, CO, and PM, respectively. Generally, these values are approximately similar to or higher than rail yard emission factors. At no electrical load, rail yard HC and CO emission rates are 7% and 14% higher, respectively, than the certification data.

**TABLE 4 Fuel-Based Emission Factors for HEP Engines for GP40 and F59PHI Locomotives for Selected Electrical Loads**

Locomotive Number, Model, Engine	Electrical Load Level <sup>a</sup>	Electrical Load (kW)	Fuel Use (g/s)	NO as NO <sub>2</sub> (g/gal)	HC <sup>b</sup> (g/gal)	CO <sup>b</sup> (g/gal)	Opacity-Based PM (g/gal)
NC1792, GP40, Cummins KTA19	None	1	4.1	39	9.6	41	1.4
	Low	14	4.8	43	3.7	19	1.6
	Medium	27	5.8	50	3.6	11	1.7
	High	53	7.7	63	3.4	9.4	1.6
NC1755, F59PHI, CAT 3412	None	1	5.2	130	12	82	2.4
	Low	8	6.0	150	8.9	60	1.6
	Medium	13	6.8	150	8.0	52	2.1
	High	26	9.2	170	6.5	38	2.1
NC1797, F59PHI, CAT 3412	None	1	5.3	110	7.9	74	2.2
	Low	na	na	na	na	na	na
	Medium	18	7.4	130	9.1	50	2.6
	High	31	9.3	140	6.3	39	2.3

NOTE: na = not available.

<sup>a</sup>For each car, all lights were turned on and air conditioning was run at daytime thermostat setting (72°F). None = only recharging of batteries; low = 1 passenger car; medium = 2 passenger cars; high = 4 passenger cars.

<sup>b</sup>Italic numbers indicate emission rates based on exhaust concentrations that are below the detection limit of gas analyzers. Detection limits for HC and CO are 13 ppm and 0.012 volume percent, respectively.



For comparison with the CAT 3412 engine, certification data for a CAT 3CPXL27 engine are used. The latter is a 27-L nonroad diesel engine with BSFC of 0.0515 gal/bhp-h. The emission rates for CAT 3CPXL27 are 123, 3.2, 32, and 4.7 g/gal for NO<sub>x</sub>, HC, CO, and PM. The rail yard emission factors for NO<sub>x</sub>, HC, and CO are higher than the certification data during load tests. The rail yard opacity-based PM emission factors are less than those of the certification data.

Fuel-based HEP NO emission rates are substantially lower than those of the much larger PMEs. The fuel-based HC emission rates were of comparable magnitude for the HEP and PMEs. For CO, the fuel-based rates ranged from 38 to 82 g/gal for the CAT 3412 engines. These rates tend to be higher than those for the PMEs. The CO emission rates for the Cummins engine are comparable with those of the PMEs. The fuel-based PM emission rates for all three HEP engines were approximately similar and appear to be higher than for the substantially larger PMEs. There is not a strong trend of fuel-based PM emission rate with respect to load.

## CONCLUSIONS

The use of the PEMS to conduct rail yard tests has been demonstrated in this work on the basis of applications to three locomotives. Rail yard tests are a relatively low-cost method for benchmarking and comparing locomotive emissions. Rail yard tests are significantly cheaper than centralized FRM tests that require sending a locomotive to a test facility, which would involve significant time out of service and lost revenue. For a local rail yard test, the locomotive is typically kept in the rail yard for a day. The total duration of preparation, testing of the PME, and testing of the HEP is approximately 6 h. For each day of testing, there are typically about 4 to 5 days of work for analyzing and reporting data.

The rail yard exhaust concentrations for NO and CO<sub>2</sub> for the PMEs were comparable with other static test results for each notch position. The HC and CO concentrations were often below the detection limit of gas analyzers. The opacity-based PM showed low concentrations compared with PM concentrations based on FRM.

The fuel-based NO<sub>x</sub> emission rates for the EMD16-645 engine were generally higher than those for the EMD12-710 engine. The fuel-based opacity was similar for both engines.

The line-haul cycle average fuel-based emission rates for NO<sub>x</sub> and CO were comparable with benchmark data based on FRMs. HC measurements using NDIR have a known bias. When this bias is considered, the measurement results are comparable with benchmark data. For PM, the semiquantitative laser light scattering measurements are useful for relative comparisons with data obtained with the same method but are not accurate in respect to absolute magnitude.

For HEP engines, the exhaust concentrations for Cummins KTA19 and CAT 3412 increased as electrical load increased. Whereas the fuel-based emission rates for NO, HC, CO, and opacity for the Cummins KTA19 engine were lower than for the benchmark data, those for CAT 3412 were higher than for the benchmark data, except opacity.

Emission factors based on PEMS measurements are useful for comparing engines. The data are reasonable and a useful benchmark to data that will be collected in future measurement campaigns. Examples of factors that will be assessed in future comparative studies include the effect of substitution of alternative fuel, such as B20 biodiesel, for ultralow sulfur diesel. Furthermore, the effect on emissions of hardware or operational modifications to the engines

will be assessed. The same or similar methodology can be applied to other locomotives. The methodology will be adapted for in-use measurement of locomotives in over-the-rail service. The advantage of this type of measurement will be to obtain real-world duty cycles that may be unique to passenger rail service, as opposed to the national average freight duty cycles provided by EPA.

## ACKNOWLEDGMENTS

Dave Krajcovic, Gary Cuthbertson, and Ed Held of Herzog Transit Services, Raleigh, North Carolina, provided valuable technical support, including fabrication of sampling ports for the engines of the tested locomotives and the operation of the locomotive engines during the tests. Sharon Mahoney of the North Carolina DOT Rail Division coordinated the scheduling of field activities. Allan Paul of the North Carolina DOT Rail Division and Dennis Pipkin of the North Carolina DOT Research and Development unit provided guidance and logistical support.

## REFERENCES

1. Bureau of Transportation Statistics. Class I Railroad Locomotive Fleet by Year Built. [http://www.bts.gov/publications/national\\_transportation\\_statistics/html/table\\_01\\_29.html](http://www.bts.gov/publications/national_transportation_statistics/html/table_01_29.html). Accessed Nov. 17, 2011.
2. *Emission Factors for Locomotives*. EPA-420-F-09-025. U.S. Environmental Protection Agency, Ann Arbor, Mich., 2009.
3. Bannikov, M. G., and J. A. Chattha. Oxides of Nitrogen (NO<sub>x</sub>) Emission Levels of Diesel Engines of Switch Locomotives. *Proc., Institution of Mechanical Engineers, Part A: Power and Energy*, Vol. 220, No. 5, 2006, pp. 449–457.
4. Markworth, V. O., S. G. Fritz, and G. R. Cataldi. The Effect of Injection Timing, Enhanced Aftercooling, and Low-Sulfur, Low-Aromatic Diesel Fuel on Locomotive Exhaust Emissions. *Journal of Engineering for Gas Turbines and Power-Transactions of the ASME*, Vol. 114, No. 3, 1992, pp. 488–495.
5. Chen, G., P. L. Flynn, S. M. Gallagher, and E. R. Dillen. Development of the Low-Emission GE-7FDL High-Power Medium-Speed Locomotive Diesel Engine. *Journal of Engineering for Gas Turbines and Power*, Vol. 125, No. 2, 2003, pp. 505–512.
6. U.S. Environmental Protection Agency. Control of Emissions of Air Pollution from Locomotive Engines and Marine Compression-Ignition Engines Less Than 30 Liters per Cylinder. Republication. Final rule. *Federal Register*, 73(126):37095–37350.
7. U.S. Environmental Protection Agency, Protection of the Environment, Control of Air Pollution from Locomotives and Locomotive Engines, Test Procedures, Analyzer Specifications, Title 40, Part 92, Subpart B, §92.109, Promulgated April 16, 1998. Amended July 13, 2005. <http://ecfr.gpoaccess.gov/cgi/t/text/text-idx?c=ecfr&sid=37ddcf709da77106546c201eb11a98ab&rgn=div8&view=text&node=40:20.0.1.1.6.2.1.9&idno=40>. Accessed April 16, 2012.
8. Fritz, S. G. *Exhaust Emissions from Two Intercity Passenger Locomotives*. American Society of Mechanical Engineers, 1994, pp. 774–783.
9. *OEM-2100 Montana System Operation Manual*. Clean Air Technologies International, Inc., Buffalo, N.Y., Nov. 2003.
10. Vojtisek-Lom, M., and J. E. Allsop. *Development of Heavy-Duty Diesel Portable, On-Board Mass Exhaust Emissions Monitoring System with NO<sub>x</sub>, CO<sub>2</sub>, and Qualitative PM Capabilities*. SAE 2001-01-3641. SAE, Warrenton, Pa., 2001.
11. Vojtisek-Lom, M., and J. T. Cobb. Vehicle Mass Emissions Measurement Using a Portable 5-Gas Exhaust Analyzer and Engine Computer Data. *Proc., Environmental Protection Agency/Air and Waste Management Association Emission Inventory Meeting*, Research Triangle Park, N.C., 1997.
12. Jimenez, J. L., J. M. Gregory, D. D. Nelson, M. S. Zahniser, and C. E. Kolb. Remote Sensing of NO and NO<sub>2</sub> Emissions from Heavy-Duty Diesel Trucks Using Tunable Diode Lasers. *Environmental Science and Technology*, Vol. 34, No. 12, 2000, pp. 2380–2387.

13. Bromberg, L., D. Cohn, A. Rabinovich, and J. Heywood. Emissions Reductions Using Hydrogen from Plasmatron Fuel Converters. *International Journal of Hydrogen Energy*, Vol. 26, No. 10, 2001, pp. 1115–1121.
14. *Environmental Technology Verification Report: Clean Air Technologies International, Inc. REMOTE On-Board Emissions Monitor*. Prepared by Battelle under a cooperative agreement with the U.S. Environmental Protection Agency, Ann Arbor, Mich., June 2003.
15. Durbin, T. D., K. Johnson, D. Cocker, J. Miller, H. Maldonado, A. Shah, C. Ensfield, C. Weaver, M. Akard, N. Harvey, J. Symon, T. Lanni, W. D. Bachalo, G. Payne, G. Smallwood, and M. Linke. Evaluation and Comparison of Portable Emissions Measurement Systems and Federal Reference Methods for Emissions from a Back-Up Generator and a Diesel Truck Operated on a Chassis Dynamometer. *Environmental Science and Technology*, Vol. 41, No. 17, 2007, pp. 6199–6204.
16. Johnson, K. C., T. D. Durbin, D. R. Cocker, J. W. Miller, R. J. Agama, N. Moynahan, and G. Nayak. On-Road Evaluation of a PEMS for Measuring Gaseous In-Use Emissions from a Heavy-Duty Diesel Vehicle. *SAE International Journal of Commercial Vehicles*, Vol. 1, No. 1, 2009, pp. 200–209.
17. Johnson, K. C., T. D. Durbin, D. R. Cocker III, W. J. Miller, D. K. Bishnu, H. Maldonado, N. Moynahan, C. Ensfield, and C. A. Laroo. On-Road Comparison of a Portable Emission Measurement System with a Mobile Reference Laboratory for a Heavy-Duty Diesel Vehicle. *Atmospheric Environment*, Vol. 43, No. 18, 2009, pp. 2877–2883.
18. Stephens, R. D., P. A. Mulawa, M. T. Giles, K. G. Kennedy, P. J. Groblicki, and S. H. Cadle. An Experimental Evaluation of Remote Sensing-Based Hydrocarbon Measurements: A Comparison to FID Measurements. *Journal of the Air and Waste Management Association*, Vol. 46, No. 2, 1996, pp. 148–158.
19. Frey, H. C., and K. Kim. Comparison of Real-World Fuel Use and Emissions for Dump Trucks Fueled with B20 Biodiesel Versus Petroleum Diesel. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 1987, Transportation Research Board of the National Academies, Washington, D.C., 2006, pp. 110–117.
20. Frey, H. C., W. J. Rasdorf, K. Kim, S. Pang, and P. Lewis. *Real-World Duty Cycles and Utilization for Construction Equipment in North Carolina*. FHWA/NC/2006-08. Prepared by North Carolina State University for the North Carolina Department of Transportation, Raleigh, 2008.
21. *Locomotive Emission Standards: Regulatory Support Document*. EPA/98-04. U.S. Environmental Protection Agency, Ann Arbor, Mich., 1998.
22. McKenna, D., K. Bhatia, R. Hesketh, C. Rowen, T. Vaughn, A. J. Marchese, G. Chipko, and S. Guran. Evaluation of Emissions and Performance of Diesel Locomotives with B20 Biodiesel Blend: Static Test Results. *Proc., 2008 ASME Rail Transportation Division Fall Technical Conference*, Chicago, Ill., 2008.
23. Poola, R. B., and R. Sekar. Reduction of NO<sub>x</sub> and Particulate Emissions by Using Oxygen-Enriched Combustion Air in a Locomotive Diesel Engine. *Journal of Engineering for Gas Turbines and Power-Transactions of the ASME*, Vol. 125, No. 2, 2003, pp. 524–533.
24. *Highlights, Emission Factors for Locomotives*. EPA420-F-97-051. U.S. Environmental Protection Agency, Ann Arbor, Mich., 1997.
25. Fritz, S. G. *Diesel Fuel Effects on Locomotive Exhaust Emissions*. SWRI Project No. 08.02062. Prepared by Southwest Research Institute for California Air Resources Board, Sacramento, Calif., Oct. 2000.
26. Fritz, S. G. *Evaluation of Biodiesel Fuel in an EMD GP38-2 Locomotive*. NREL/SR-510-33436. Prepared by Southwest Research Institute for the National Renewable Energy Laboratory, Golden, Colo., 2004.
27. Barth, M., T. Younglove, and G. Scora. *Development of a Heavy-Duty Diesel Modal Emissions and Fuel Consumption Model*. UCB-ITS-PRR-2005-1. Prepared by University of California, Riverside, for California PATH Program, Jan. 2005.

---

*The contents of this paper reflect the views of the authors and not necessarily the views of North Carolina State University or the official views or policies of the North Carolina Department of Transportation, FHWA, or the Institute for Transportation Research and Education. The authors are responsible for the facts and accuracy of the data presented. This report does not constitute a standard, specification, or regulation.*

*The Passenger Rail Equipment and Systems Integration Committee peer-reviewed this paper.*

# Slab Track Mass-Spring System

Mirjana Tomicic-Torlakovic, Miodrag Budisa, and Vidan Radjen

**Development of a ballastless slab track structure with a mass-spring system began about 20 years ago, and different kinds of these track systems have been successfully in use for urban transportation systems. With these systems the elastic elements are inserted under the slabs to provide protection from bothersome vibration and noise. The paper explains the principal aspects of calculation verifications by means of a simulation model for the selected type of mass-spring track system.**

Rail traffic is an integral part of the public transportation systems in the central areas of many cities. Proximity to neighboring buildings, the sensitivity of the population, and the necessity of sharing the route with motor traffic are factors in track design and construction.

The vibrations from rail traffic are caused mainly by the rolling of the running wheel on the rails (direct structure-borne noise); the vibrations propagate via the superstructure and the tunnel through the ground into nearby buildings and can cause audible secondary airborne noise. Railway tracks in an urban environment usually have to fulfill certain requirements concerning the emission of noise and vibration. The decisive range of vibration is 5 to 20 Hz and for structure-borne air noise the range is 40 to 80 Hz (1).

Comparisons between different possibilities for reducing noise and vibration showed that measures at the railway superstructure itself are usually preferable from a technical and economic point of view. A wide variety of measures are available, and there are effective tools for reducing structure-borne noise at the source. These measures include the use of highly elastic pads for rail fasteners, ballast, sleeper mats, and elastic supports for slab tracks of so-called mass-spring systems. The decisive parameter for vibration absorption is the natural frequency (eigen frequency) of the selected superstructure system.

Some examples of successfully completed mass-spring system projects for standard gauge railways are the Römerberg Tunnel, Zammer Tunnel, Arlberg Tunnel, New Lainz Tunnel, and Sittenberg Tunnel, in Austria; the Tiergarten Tunnel (Berlin North–South) in Germany; the Zimmerberg Tunnel (Swiss national railway) in Switzerland; Rome–Fiumicino, Udine–Tarvisio, Milan–Saronno, and Catania (Italian national railway), in Italy; the Channel Tunnel Rail Link (Network Rail) between England and France; and Brussels (Belgian national railway) in Belgium. Examples for tram lines include Augsburg, Germany; Barcelona, Spain; Berlin; Bern, Switzerland; Essen, Germany; Florence, Italy; Geneva; Graz,

Austria; Milan, Italy; Munich, Germany; Nice, France; Seville, Spain; Stuttgart, Germany; Valencia, Spain; and Vienna, Austria. Examples for underground and rapid transit lines include Athens, Greece; Berlin; Buenos Aires, Argentina; Dortmund, Germany; Hong Kong; Krakow, Poland; Milan, Italy; New York; Sao Paulo, Brazil; and Zurich, Switzerland.

## TYPES OF MASS-SPRING SYSTEMS

Mass-spring systems are used in applications in which the isolation demands in regard to structure-borne noise are very high (Figure 1). In recent decades, a wide range of mass-spring systems have been developed. There are systems that use in situ concrete or prefabricated concrete components or their combination, with or without a ballast bed. The type of construction chosen is a basic factor in the design of elastic supports for mass-spring systems.

Figure 2 shows three different types of such systems (2):

- Full surface layer,
- Linear support, and
- Discrete bearings.

In all types of such systems, elastic elements make the following possible (2):

- Reducing the loads on the superstructure, the substructure, and the subgrade;
- Reducing the wear on rails and wheels;
- Increasing track elasticity;
- Protecting the environment against vibrations and structure-borne noise;
- Making construction fast; and
- Keeping track maintenance costs very low.

## MASS-SPRING SYSTEMS WITH DISCRETE BEARINGS

Discrete bearings are predetermined by the form of the track slab or track trough sections. These slabs can be prefabricated or cast on site.

When the slabs are cast on site, an insulation sheet is placed on the foundation to prevent it from bonding, so that the slab can be raised later. The slab is lifted after hardening, and then the bearings are inserted through the installation holes in the slab sections. If the longitudinal sides of the slab are free, it is possible to lift the slab with jacks that can be applied from the sides (Figure 3).

The joints between the slabs are connected with plastic-coated pins, which will be cemented with mortar or with resin in a dovetail pattern. The bearings close to the joints will be 1.5 times stiffer than the standard bearings.

M. Tomicic-Torlakovic, Civil Engineering Faculty, University of Belgrade, Bulevar Kralja Aleksandra 73, Belgrade 11000, Serbia. M. Budisa, Alpha Rail OpenTrack Consulting Inc., USA, 2766 South Logan Street, Suite 100, Englewood, CO 80113. V. Radjen, Inter-Kop, Bulevar Mihajla Pupina 115, Belgrade 11000, Serbia. Corresponding author: M. Budisa, opentrack.USA@gmail.com.

*Transportation Research Record: Journal of the Transportation Research Board*, No. 2289, Transportation Research Board of the National Academies, Washington, D.C., 2012, pp. 64–69.  
DOI: 10.3141/2289-09

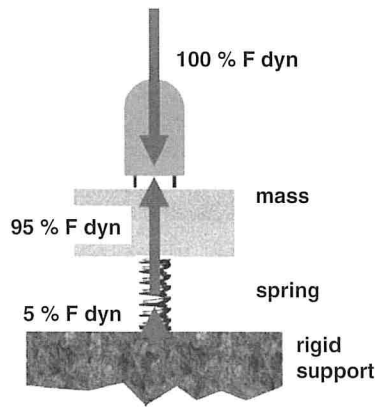


FIGURE 1 Damping effect of mass-spring system.

In the transition zones between the mass-spring system and the conventional slab track, bearings with a higher degree of stiffness are installed to reduce track depression, prevent excessive track fatigue, and reduce the risk of cracking (3).

Because of the relatively low surface area of the supports, special attention must be paid to the horizontal forces arising from train operations. To limit the deflections as required, the perfect balance between the shear modulus, elasticity, support thickness, and surface area of the support must be found.

By using individual bearings, the lowest tuning frequencies can be achieved, depending on the mass of the superstructure, with the following well-known formula:

$$f = 1/2\pi(k/m)^{1/2} \text{ (Hz)} \quad (1)$$

where  $k$  equals the elasticity of slab bearings per 1 m of the slab (N/m) and  $m$  equals the mass of the slab per 1 m (kg).

The natural (eigen) frequency is between 5 and 10 Hz. A low eigen frequency of less than about 20 Hz is required. This allows the maximum isolation level against structure-borne noise of up to 30 dB (3).

There is a discrepancy between the required natural structure frequency and the rail deflection: the softer the bearing material, the lower the natural frequency and the higher the deflection will be.

A new generation of high-performance elastomers (for example, Sylomer by Getzner) serve as the materials for elastic supports in all mass-spring systems. The elastomers offer many advantages when used as elastic bearings for slab track and ballast mats, including the following (3):

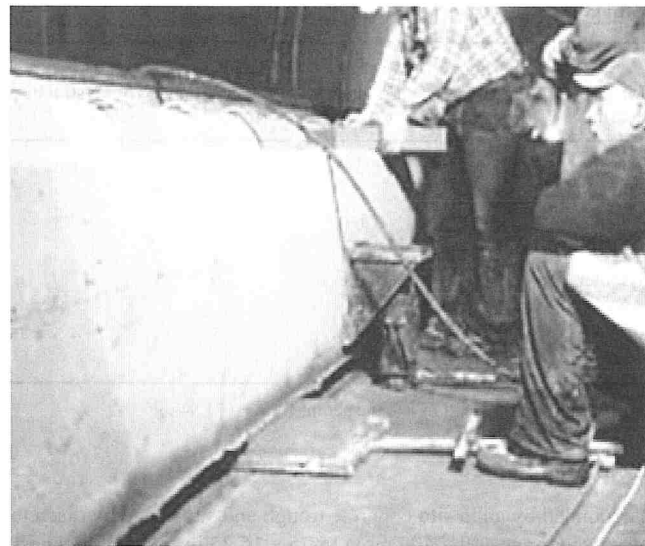


FIGURE 3 Installation of discrete bearings.

- Reliable, homogeneous, and durable elastic properties;
- Resistance to short-term, extreme overloading;
- Ease of use for compensating construction tolerances; and
- Adaptability to all applications with variations in the density of the material, the thickness, and the surface area of the support.

The great number of Sylomer bearings installed in mass-spring systems justifies previous statements (Table 1) (3).

Natural rubber is a widely used material for discrete bearings [for example, Trackelast by Tiflex (4)] mainly because of its unbeatable dynamic characteristics and long proven service record on bridges.

## MODELING HEAVY MASS-SPRING SYSTEMS

The model of this type of track superstructure system is designed as a so-called "heavy mass-spring" system (Figure 4) for open track sections of light rail public transportation systems in cities (5). It means that the slab is thick enough, and consequently massive enough, that with Formula 1, the system has a low natural (eigen) frequency. The following is an example calculation of a heavy mass-spring system.

The structure consists of a prefabricated concrete track slab with dimensions of  $240 \times 390 \times 50$  cm ( $94.5 \times 153.5 \times 19.7$  in.),

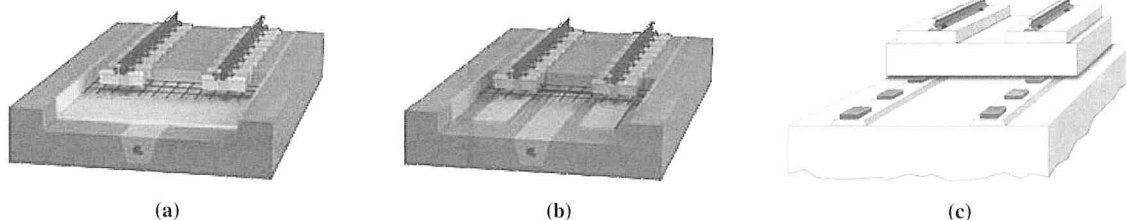


FIGURE 2 Types of mass-spring systems.

TABLE 1 Types of Sylodyn and Sylomer Bearings

Type of Bearing	Bearing Thickness (mm)	Natural Frequency (Hz)	Track Deflection (mm)	Mass of Superstructure (t/m)
Sylodyn N 70	70	5.2	9.2	11.0
Sylodyn N 70	59	6.0	7.0	7.5
Sylomer S 50	50	8.0	3.9	7.0
Sylomer S 50	50	10.0	2.5	4.5

NOTE: 1 mm = 0.039 in.; t/m = tonne per m; 1 t/m = 671.94 lb/ft.

surrounded by an in situ concrete trough and supported by discrete elastic bearings  $400 \times 400$  mm ( $15.75 \times 15.75$  in.) with a thickness of 40 mm (1.57 in.) (Figure 5). The static elasticity coefficient is  $30.7 \text{ N/m}^3$  ( $0.196 \text{ lbf/ft}^3$ ), and the density is  $400 \text{ kg/m}^3$  ( $24.953 \text{ lb/ft}^3$ ).

The input for track parameters is

- Type of rail: 49E1 (49.43 kg/m or 99.65 lb/yd),
- Fasteners spacing: 65 cm (25.59 in.),
- Rail pad elasticity:  $200 \text{ kN/cm}$  ( $570.58 \text{ lbf/in.}$ ),
- Slab track natural (eigen) frequency: mass of slab per meter =  $2.4 \text{ m} \times 0.5 \text{ m} \times 2,400 \text{ kg/m}^3 = 2,880 \text{ kg}$  ( $94.49 \text{ in.} \times 19.69 \text{ in.} \times 3.408 \text{ psi} = 6,340.61 \text{ lb}$ ),
- Elasticity of slab bearings per meter of the slab (two bearings every 2 m):  $30.7 \times 0.4 \times 0.4 = 0.49 \times 10^7 \text{ N/m}$  ( $336,140 \text{ lbf/ft}$ ), and
- Natural (eigen) frequency using Formula 1:  $f = 1/2\pi(0.49 \times 10^7/2,880)^{1/2} = 7 \text{ Hz}$ .

The properties for superstructure (slab track) and substructure elements are illustrated in Table 2.

The competent load is the electric locomotive passenger [German high-speed train, or the InterCity Express (ICE)] train with an axle load of  $196 \text{ kN}$  ( $44,062.6 \text{ lbf}$ ) and an axle distance of  $3.0 \text{ m}$  ( $9.84 \text{ ft}$ ). The rail distributes this load over the solid track slab via the rail supports.

The elastic line of the rail and rail supporting forces acting on the slab (Figure 6) are calculated by using Zimmerman's theory

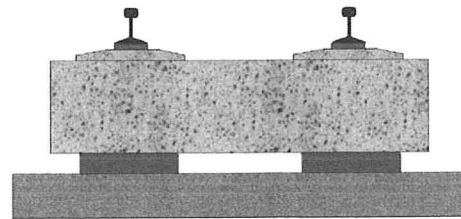


FIGURE 4 Heavy mass-spring system.

for an elastically supported beam, with Vincler's hypothesis, as follows (6):

Rail deflection is

$$y = \frac{k}{2U} \sum Q_i \eta_i \quad (\text{mm}) \quad (2)$$

where

$Q_i$  = vehicle load forces,

$\eta_i = \frac{\sin(k \cdot x_i) + \cos(k \cdot x_i)}{e^{(k \cdot x_i)}} = \text{influence coefficient of axle distance } x_i,$

$x_i$  = distance between axles (m),

$k = \sqrt[4]{\frac{U}{4EI_x}} = \text{coefficient of relative stiffness of the track supports using rail,}$

$U = D/L = \text{elasticity modulus of track,}$

$D = \text{track support stiffness,}$

$L = \text{support distance, and}$

$EI_x = \text{bending stiffness of rail.}$

Supporting forces are

$$S = \frac{kL}{2} \sum Q_i \eta_i \quad (\text{N}) \quad (3)$$

The forces and the deformations of the slab track structure are calculated by means of the TOWER program, which uses the finite element method (7).

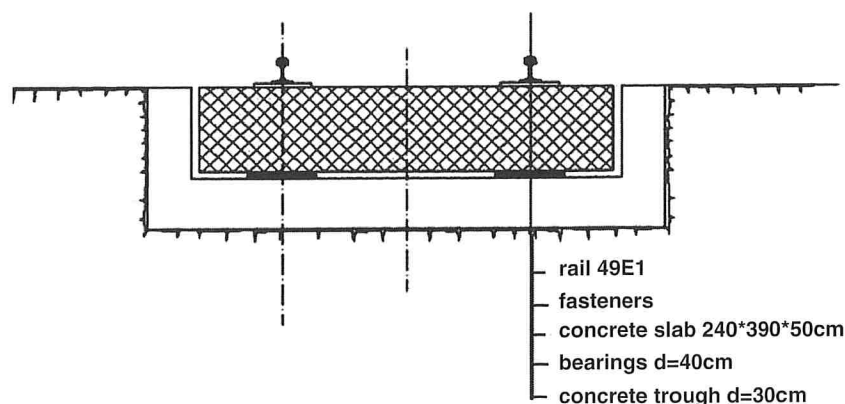
FIGURE 5 Example of characteristics of heavy mass-spring system ( $d$  = thickness).

TABLE 2 Properties of Slab Track Superstructure and Substructure Elements

Superstructure Property	$E'$ (MN/m <sup>2</sup> )	Density (t/m <sup>3</sup> ) <sup>b</sup>	Poisson's Ratio	Coefficient of Damping (%)
Concrete slab (B30)	31,500	2.5	0.2	1.0
HSL (B10)	22,000	2.3	0.1	1.0
FPL ( $E_{vz} = 120$ MN/m <sup>2</sup> ) <sup>c</sup>	435	1.9	0.3	1.0
Soil ( $E_{vz} = 40$ MN/m <sup>2</sup> ) <sup>c</sup>	180	1.7	0.3	2.5

NOTE: HSL = hydraulically stabilizing layer; FPL = frost protection layer; 1 MN/m<sup>2</sup> = 145.16 psi; MN = meganewton.

<sup>a</sup> $E$  = modulus of elasticity.

<sup>b</sup>1 t/m<sup>3</sup> = 62.43 lb/ft<sup>3</sup>.

<sup>c</sup> $E_{vz}$  = modulus of deformability.

The settlements of the substructure are obtained from Schlicher's formula (5):

$$s = \frac{(1 - \nu_0^2)P}{E_0 F} B_\alpha \quad (\text{mm}) \quad (4)$$

where

- $\nu_0$  = substructure Poisson's coefficient,
- $E_0$  = substructure elasticity modulus,
- $P$  = slab load (N),
- $F$  = track structure supporting surface (mm<sup>2</sup>),
- $B$  = width of track structure (m), and
- $\alpha$  = coefficient depending on shape of structure surface and position of point where settlement is determined.

The substructure stiffness is

$$k = \frac{P}{s} = \frac{98 \cdot 8}{0.0054} = 145,185 \text{ kN/m} \quad (5)$$

The stiffness of the members that simulate the slab bearings is

$$k_{zi} = cF_{ppi} \quad (6)$$

where  $c$  equals 30.7 N/cm<sup>3</sup> equals elasticity coefficient of Sylomer, and  $F_{ppi}$  equals the slab area that corresponds to one member.

The model consists of five track slabs, of the shape shown in Figure 4, elastically supported on the substructure with a stiffness value defined by Formula 5. An illustration of the three-dimensional computational model is presented in Figure 7. The loads for the rail supporting forces calculated for the ICE train are represented in Figure 5. The position of the load on the slabs is shown in Figure 8 (5). Figure 9 shows the diagram of track slab deflections and the deformed shape of the model (5).

Further, with the assumed properties for superstructure (slab track) and substructure elements, the time history for the vertical slab track displacements in the competent joint (with maximum dynamic deflection) of the model is shown in Figure 10.

A greater dynamic deflection is expected for lower velocities (for example, 0.7 mm or 0.03 in. for 50 km/h or 31 mph as required in the central zones of cities, where the rail traffic shares a route with motor traffic) (8). For higher velocities, the deflections of the slab track would be even smaller. The main reason for this is that with an increase in velocity, the static effect of the load decreases. Also, the higher excitation frequencies caused by the higher velocities require the greater mass of the structure and the soil.

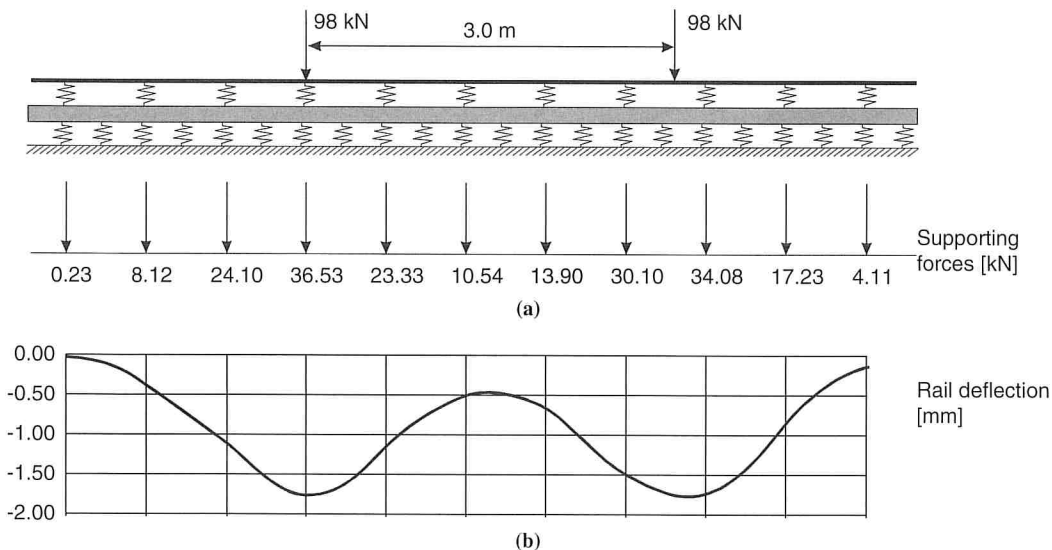


FIGURE 6 Rail (a) supporting forces and (b) deflections for ICE electric locomotive.



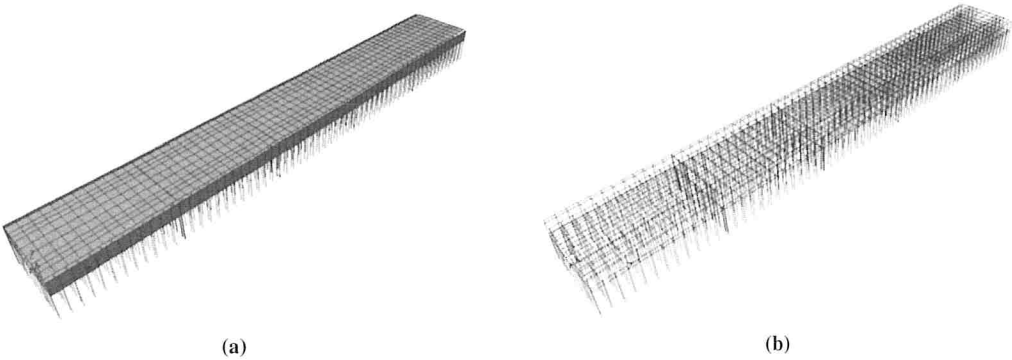


FIGURE 7 Computational model in three-dimensional views: (a) solid elements and (b) beam elements.

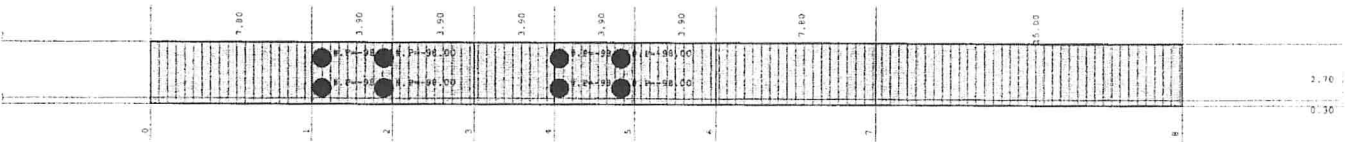


FIGURE 8 Position of load on the slabs.

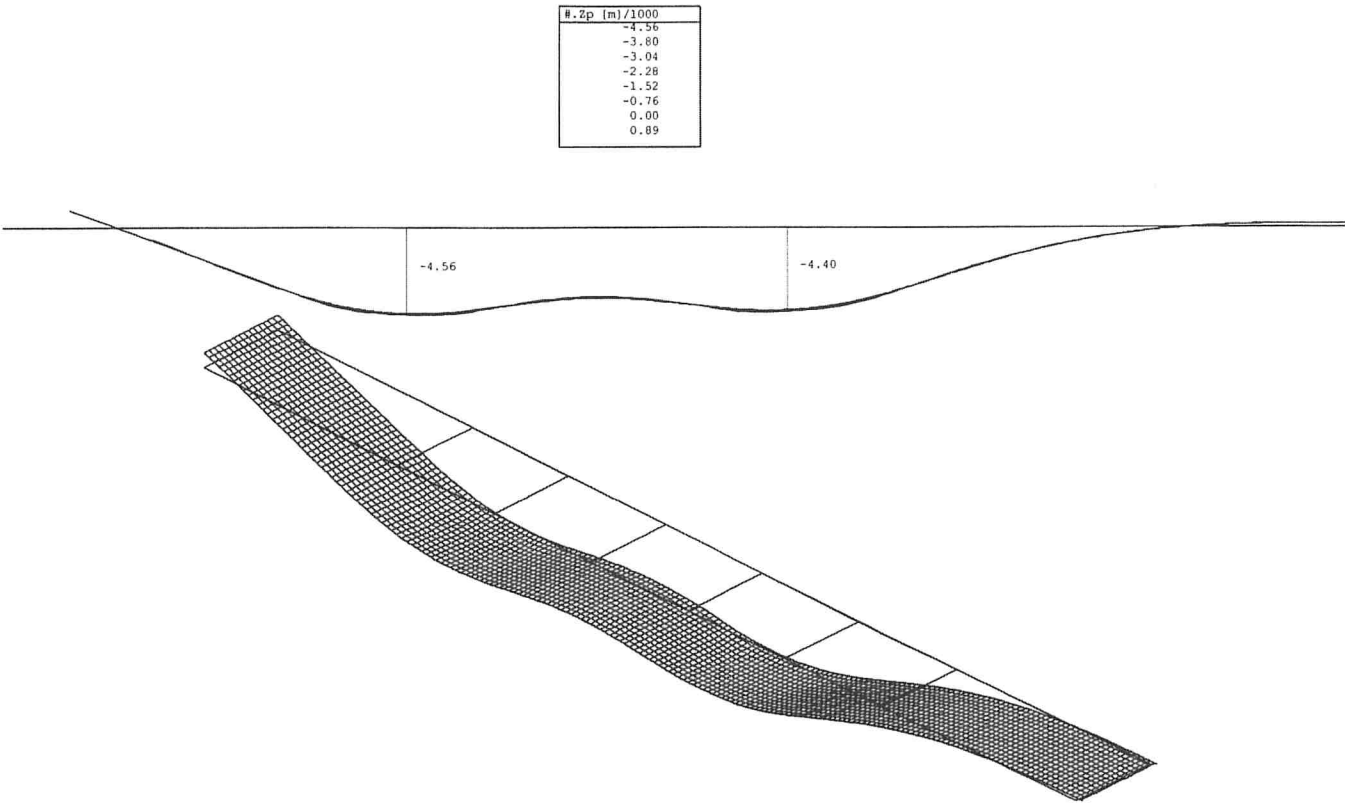


FIGURE 9 Track slab deflections and deformed shape of the model.

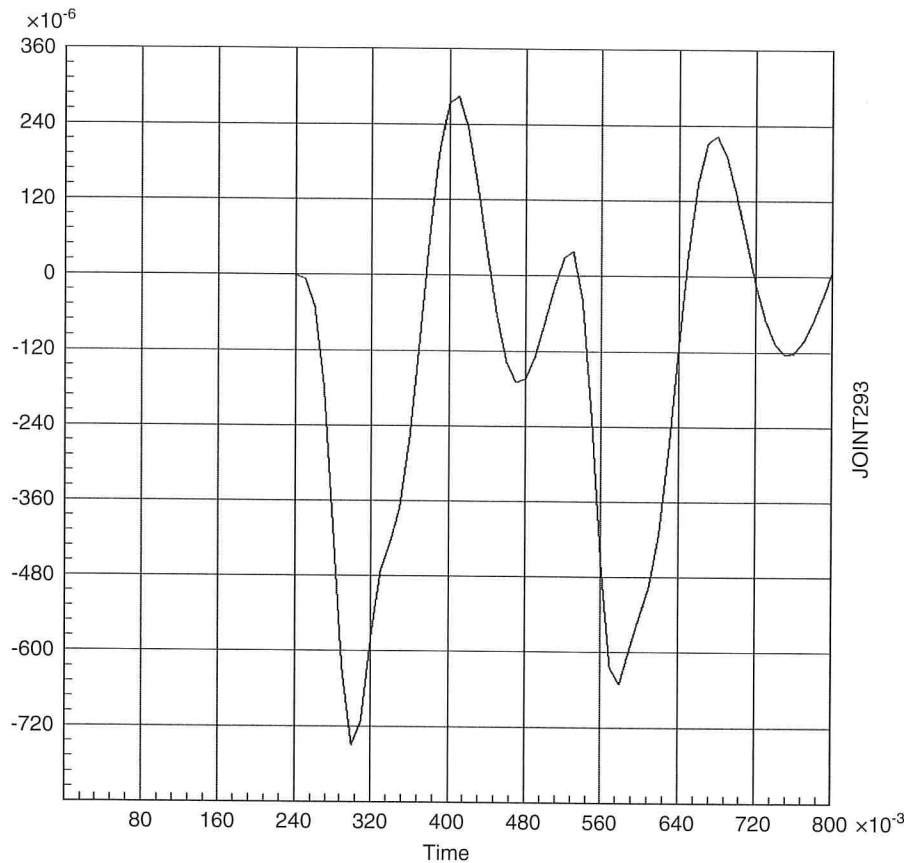


FIGURE 10 Time history for vertical displacement in competent joint.

## CONCLUSION

Although the track slab mass-spring system is fairly expensive, it is used because it completely fulfills the special requirements of urban areas, especially the slab deflection limits and isolation from vibration and noise at the source. Depending on the elastic element materials used, it is possible to influence the amount of slab stress and strain state of the mass-spring system within wide limits. In the example presented, the maximum static deflection of 4.56 mm (0.18 in.) is acceptable because it agrees with the measured deflection (6, 9).

## REFERENCES

1. Lenz, U., H. Stank, and H. J. Stummeyer. Dimensionierung von Masse-Feder-Systemen für Eisenbahn (in German). *Eisenbahningenieur*, March 2007, pp. 12–18.
2. Getzner Werkstoffe GmbH. *Elastic Supports for Slab Tracks and Ballast Troughs—“Mass-Spring” System*. <http://www.getzner.com/en/downloads/brochures>. Accessed 2003.
3. Getzner Werkstoffe GmbH. *Inform aktuell*, 2nd (2000), 3rd (2000), and 2nd (2001) eds.
4. Tiflex. *Trackelast-Bearings Brochure*. [http://www.figlex.co.uk/track\\_home/fst/fst/html](http://www.figlex.co.uk/track_home/fst/fst/html). Accessed 2010.
5. Radjen, V. *Track Superstructure of “Heavy Mass-Spring” System*. Diploma work. Civil Engineering Faculty University of Belgrade, Serbia, 2010.
6. Tomicic-Torlakovic, M., and S. Lelovic. Static Calculation of the Track “Mass-Spring” System. *Bulletins for Applied and Computer Mathematics*, No. 2148. Pannonian Applied Mathematical Meetings, Balaton, Hungary, 2003, pp. 549–555.
7. *TOWER User's Manual*. Belgrade, Serbia, 2007.
8. Verbic, B., G. Schmid, H. D. Köpper, and H. Best. Investigating the Dynamic Behavior of Rigid Track. *Railway Gazette International*, Sept. 1997, pp. 583–586.
9. Tomicic-Torlakovic, M., and L. Puzavac. Contribution to Calculation of the Slab Track “Mass-Spring” System. *ZEL (Rail Symposium)*, Ziline, Slovakia, 2003.

*The Railroad Track Structure System Design Committee peer-reviewed this paper.*

# Comparison of Magnitude of Actions on Track in High-Speed and Heavy-Haul Railroads

## Influence of Resilient Fastenings

Konstantinos Giannakos

In modern high-speed railways the necessity for achieving low life-cycle costs has led to a well-compacted—almost undeflected—substructure. Applications of excellent quality substructure are encountered in the French network with the operation of high-speed trains (TGV in French) and also in research projects of the International Union of Railways. The countervailing need for increased elasticity in the track's structure is satisfied by the development of the highly resilient fastenings with soft pads. There is a long-standing discussion about whether these soft pads promote a sufficient life cycle for heavy-haul railways as well, as in the case of the U.S. railways. In this paper a parametric investigation using a theoretical approach is performed to estimate the actions on track support points (ties) in high-speed and heavy-haul lines with four methods found in the international literature; parameter values, mainly coefficients of stiffness, derived from measurements on the track are used. Because the life cycle of the pads and the clips of the fastenings are dependent on the actions exerted on each tie per passing axle and the number of axle passes, this investigation leads to significant ascertainment for comparing the actions on the track in both cases.

The railway track infrastructure appears to be the equivalent of a flexible pavement structure, relatively simple, following the rules and specifications of proper design. It seems to have changed little in the past 250 years according to one of the patriarchs of railroad engineering in the United States, William Hay (1). According to Hay, "track development came mostly through trial and error" (1). This situation changed with the development of high-speed railways. In modern high-speed railways the necessity for achieving very low life-cycle costs and, consequently, low operating and maintenance costs, led to a very well compacted substructure, almost undeflected, with 100% modified Proctor or 105% Proctor compaction (2). Applications of excellent quality substructure are encountered in the French network with the operation of TGV (high-speed train, in French) and also in research projects of the International Union of Railways [see *Optimum Adaptation of the Conventional Track to Future Traffic* (3)]. The need for minimal maintenance and undeflected support to avoid permanent deformations leads to very stiff construction. The modern structure of the track's superstructure was finally fixed in the past 50 years (4).

The railway track is a multilayered structure consisting of a vertical succession of various materials or layers of materials that define the

final position of the rail running table as well as the properties of the track itself, as it reacts to the action that is created from the motion of the railway vehicle. Each material or layer that constitutes the line can be simulated by a combination of a spring with spring constant  $k_i$  and a dashpot with damping coefficient  $c_i$ . The weak links in this multilayered construction are the ballast and the substructure. These develop residual deformations proportionally related to the deterioration of the geometry of the track because the permanent deformation of the track is a percentage of the total deflection resulting from the train circulation.

According to Winkler's theoretical analysis of the track as a continuous beam on elastic foundation, a low-stiffness track is advantageous in that it permits deflections, distributes the loads to the adjacent ties, and thus reduces the stress and strain on the ties (support points) (5). The deflection of the track as an infinite beam on elastic foundation must be high enough to distribute the acting loads to a longer section of the track and thus to reduce the reacting force on the ties. With the evolution of specifications toward an undeflected substructure, this amount of deflection can be provided by a resilient fastening and its rail pad [see also FRA (6)]. This system should be compatible with the clip so that the toe load remains more than 7 kN. Three out of the five layers that constitute the track structure, namely, the rail, tie, and ballast, contribute only 6% to 10% to the total static track stiffness for frequencies up to 50 to 70 Hz. But even for higher frequencies that is valid because the total dynamic track stiffness  $h_{TR}$ , as in Equation 12, is a function of the total static stiffness of the track. The total track stiffness is affected mainly by the static stiffness coefficients of the pad and of the substructure. The resilient pad operates elastically with no permanent deformation and gives the appropriate elasticity to the track's structure, reducing the actions and reactions on the ties.

There is a general discussion in railway engineering circles and among academics that in heavy-haul railroads [wheel loads of 17.69 tonnes (t) or 39,000 lb and maximum speed of 60 mph or 96.6 km/h] the resilient pads undertake much higher loads per tie than in the high-speed lines of mixed traffic (wheel loads 11.25 t or 24,800 lb and  $V \geq 250$  km/h or 155.34 mph). All theoretical solutions available in the international literature are based on exactly the same theoretical approach (based on Winkler's theory, also adopted by Zimmermann). According to Eisenmann and Mattner, the theoretical calculation of actions on a railway track yields results equal to the average of the measurements on the track under operation (7). This paper presents a comparison of analysis results by using the Giannakos method and methods found in the American, French, and German literature (8).

---

Department of Civil Engineering, University of Thessaly, Volos, 108 Neoreion Str., Piraeus 18534, Greece. k.giannakos@on.gr.

*Transportation Research Record: Journal of the Transportation Research Board*, No. 2289, Transportation Research Board of the National Academies, Washington, D.C., 2012, pp. 70–77.  
DOI: 10.3141/2289-10

## THEORETICAL ANALYSIS OF TRACK AS A BEAM ON ELASTIC FOUNDATION AND MEASUREMENTS ON TRACK

The theoretical analysis is based mainly on Winkler's theory of an infinite beam on an elastic foundation. In the European literature the theory is also referred to as Zimmermann's theory (9). Apparently having in mind the tests on a track under operation performed by the European Railways and the International Union of Railways (U.I.C.), Eisenmann has stated since 1988 that the methods based on Zimmermann's theory yield results corresponding to the average of measured track values for the loading and stressing of the track as well as the track's deflection (4). Consequently, the level of maximum values is dependent on the possibility of exceedance.

In the American literature this analysis is described by Hay (1), the American Railway Engineering and Maintenance-of-Way Association (AREMA) (10), and Selig and Waters (11). All these references are based on the same theoretical analysis of a continuous beam on an elastic foundation.  $\beta$  represents the elastic length  $L$ , as used in Equation 4. The maximum deflection and moment are (10)

$$y_{\max} = w(0) = \frac{\beta \cdot Q_{\text{total}}}{2 \cdot k} \quad (1)$$

$$M_{\max} = M(0) = \frac{Q_{\text{total}}}{4 \cdot \beta} \quad (2)$$

where  $Q_{\text{total}}$  is the total load acting on track, static and dynamic, without any relation to probability of occurrence, and  $k$  in  $\text{lb/in.}^2$  is the rail support modulus derived by the relation  $p = k \cdot w = k \cdot y$ . According to AREMA,

the rail support modulus "k" ("u" is also commonly utilized in lieu of "k") as the load (in pounds) that causes a one-inch vertical rail deflection per linear inch of track. The value of "k" depends on the quality of the ties, the tie spacing, tie dimensions (foot-print and flexibility), the ballast (cleanliness, moisture content, temperature, compaction, and depth), and the subgrade (load carrying capacity and uniformity). (10)

In the European literature (French, Greek, and German) the magnitude of the "rail support modulus" or "total track stiffness (static)" is  $\text{kN/mm}$ , and it is represented by either  $\rho_{\text{total}}$  or  $c$  and can be easily derived (12):

$$k = \frac{\rho_{\text{total}}}{\ell} \Rightarrow \rho_{\text{total}} = \frac{k}{\ell} \quad (3)$$

where  $\ell$  equals the distance between ties and  $k$  rail support modulus according to AREMA.

In the European literature the elastic length  $L$  is defined as being equated to  $\beta$  of the American literature through the following:

$$\beta = \sqrt[4]{\frac{k}{4 \cdot E \cdot J}} = \sqrt[4]{\frac{\rho_{\text{total}}}{4 \cdot E \cdot J \cdot \ell}} = \frac{1}{L} \quad (4)$$

where  $E$  is the modulus of elasticity of rail and  $J$  is the moment of inertia of rail.

The influence curve for  $w$  (that is, for deflection  $y$ ) given in AREMA is used to determine the largest value  $p_{\max}$  and the maximum rail seat load  $F_{\max}$  on an individual tie (also reaction per tie/support point  $R_{\max}$ ) (10):

$$\begin{aligned} F_{\max} = R_{\max} = p_{\max} \cdot \ell &= k \cdot w_{\max} \cdot \ell = k \cdot y_{\max} \cdot \ell = k \cdot \frac{\beta \cdot Q_{\text{total}}}{2 \cdot k} \cdot \ell \\ &= \sqrt[4]{\frac{\rho_{\text{total}}}{4EJ\ell}} \cdot \frac{Q_{\text{total}} \cdot \ell}{2} = \frac{1}{2\sqrt{2}} \cdot \sqrt[4]{\frac{\rho_{\text{total}} \cdot \ell^3}{EJ}} \cdot Q_{\text{total}} = \bar{A}_{\text{stat}} \cdot Q_{\text{total}} \end{aligned} \quad (5)$$

where  $\bar{A}_{\text{stat}}$  is the same as in equations in the European literature and is given by

$$\bar{A}_{\text{stat}} = \frac{1}{2\sqrt{2}} \cdot \sqrt[4]{\frac{\ell^3 \cdot \rho_{\text{total}}}{E \cdot J}} \quad (6)$$

The total load (static and dynamic) acting on the track,  $Q_{\text{total}}$ , is not dependent on a probability of occurrence but on an impact factor  $\theta$  (1, 10):

$$\theta = \frac{D_{33} \cdot V}{D_{\text{wheel}} \cdot 100} \quad (7)$$

where

$D_{33}$  = diameter of a 33-in. wheel (in.),

$D_{\text{wheel}}$  = diameter of wheel of vehicle examined (in.), and

$V$  = speed (mph).

The total load is

$$Q_{\text{total}} = Q_{\text{wheel}} (1 + \theta) \quad (8)$$

In the German literature the reaction  $R_{\max}$  per tie is also given by Equation 5. However, the maximum  $Q_{\text{total}}$  is dependent on the probability of occurrence, and for 99.7% probability  $Q_{\text{total}}$  is given by

$$Q_{\text{total}} = Q_{\text{wheel}} \left( 1 + 0.9 \cdot \left( 1 + \frac{V - 60}{140} \right) \right) \quad (9)$$

Following an investigation program in the Greek network, after the appearance of extensive cracks in concrete ties laid on a track that exceeded 60% (see Figures 1 and 2) resulting from underestimation of actions, a method was developed that is able to predict the observed conditions on the track (12). The actions on the track panel are calculated through the following equation covering a probability of occurrence of 99.7% [for the probability of exceedance see Harrison and Ahlbeck (13)]:

$$R_{\max} = (Q_{\text{wheel}} + Q_{\alpha}) \cdot \bar{A}_{\text{dynam}} + 3 \cdot \sqrt{\sigma(\Delta R_{\text{NSM}})^2 + \sigma(\Delta R_{\text{SM}})^2} \quad (10)$$

where

$Q_{\text{wheel}}$  = static wheel load;

$Q_{\alpha}$  = load due to cant deficiency;

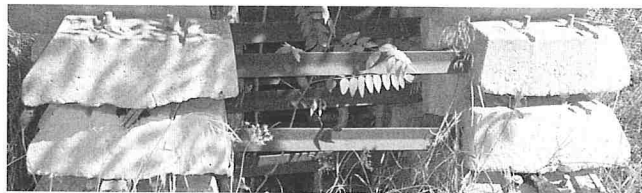


FIGURE 1 Side view of U3-type twin-block concrete ties, cracked and retired from track.

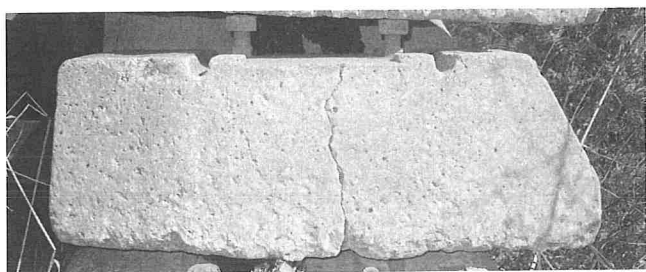


FIGURE 2 Detail from cracked block of U3-type twin-block concrete tie depicting typical crack at middle of block and its direction from lower tie seating surface on ballast toward upper surface of tie at position of rail seat.

$\bar{A}_{\text{dynam}}$  = dynamic coefficient of sleeper's reaction, 3 coefficient of dynamic load for a 99.7% probability of occurrence, or 3 times the standard deviation of total dynamic component of load  $Q$ ;

$\sigma(\Delta R_{\text{NSM}})$  = standard deviation of dynamic load due to non-suspended masses;

$\sigma(\Delta R_{\text{SM}})$  = standard deviation of dynamic load due to suspended masses [for details see Giannakos and Loizos (14)]; and

$$\bar{A}_{\text{dynam}} = \frac{1}{2\sqrt{2}} \cdot \sqrt{\frac{\ell^3 \cdot h_{\text{TR}}}{E \cdot J}} \quad (11)$$

where the total dynamic stiffness of the track  $h_{\text{TR}}$  is given by

$$h_{\text{TR}} = \frac{1}{2 \cdot \sqrt{2}} \cdot \sqrt{E \cdot J \cdot \frac{\rho_{\text{total}}}{\ell}} \quad (12)$$

The results of this method are in agreement with observations on tracks under operation. There is also a method cited in the French literature [see Alias (15) and Prud'homme and Eriau (16)] covering

a 95.5% probability of occurrence and distributing the total acting load with reaction per tie  $1.35 \cdot \bar{A}_{\text{stat}} Q_{\text{total}}$  as follows:

$$R_{\text{max}} = \bar{A}_{\text{stat}} \cdot 1.35 \cdot \left[ \frac{Q_{\text{wheel}} \cdot \left( 1 + \frac{Q_{\alpha}}{Q_{\text{wheel}}} \right)}{+ 2 \cdot \sqrt{\sigma(\Delta Q_{\text{NSM}})^2 + \sigma(\Delta Q_{\text{SM}})^2}} \right] \quad (13)$$

In the theoretical calculations above, the total static stiffness of the track plays a key role: the more elastic the track, the less the ties are stressed. It is therefore evident that resilient fastenings play a key role in the distribution of loads on the track, the stressing of the ties, and eventually in the life cycle of the track. The influence curves for  $M$ ,  $R$ ,  $y$ , and  $q$  ["intensity of pressure against rail" according to Hay (1) or uniform reaction and action along the track] are given by the following equation (10):

$$\eta(x) = e^{-\beta x} [\cos(\beta x) + \sin(\beta x)] \quad \mu(x) = e^{-\beta x} [\cos(\beta x) - \sin(\beta x)] \quad (14)$$

where

$\eta(x)$  = influence line of reaction  $R$ , deflection  $y$ , and unitary reaction  $q$  (per unit length of track);

$\mu(x)$  = influence line of moment  $M$ ;

$x$  = length of track in mm with zero (0) value just at point of application of load  $Q$ ; and

$\beta$  = value given by Equation 4.

The influence curves for track length from  $-8\pi/4$  to  $+8\pi/4$  are depicted in Figure 3.

German railways measure the stiffness  $\rho_i$  of different materials and layers of railway tracks, and the results of these measurements are also used by the French railways (14, 17, 18). Measurements of the stiffness of a track on site can be found in the American literature (19, 20). However, because the theoretical analysis should consider the surrounding curve for the most adverse conditions, the German railway measurements are more appropriate to estimate the

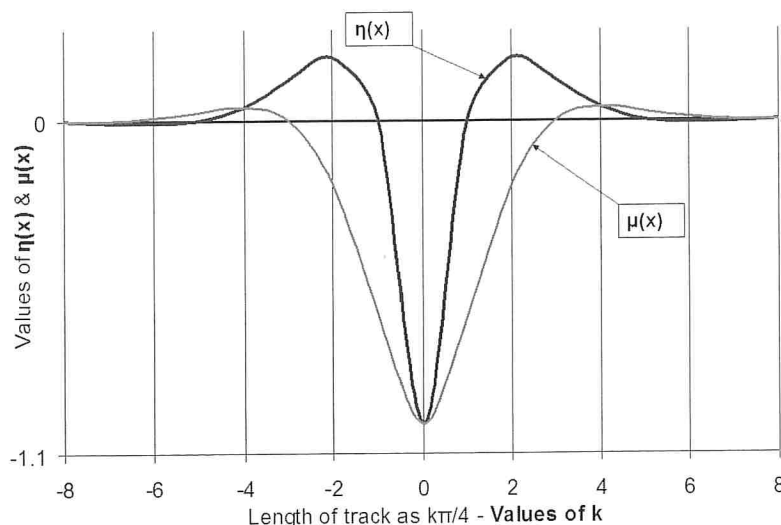


FIGURE 3 Influence curves  $\mu(x)$  and  $\eta(x)$ .



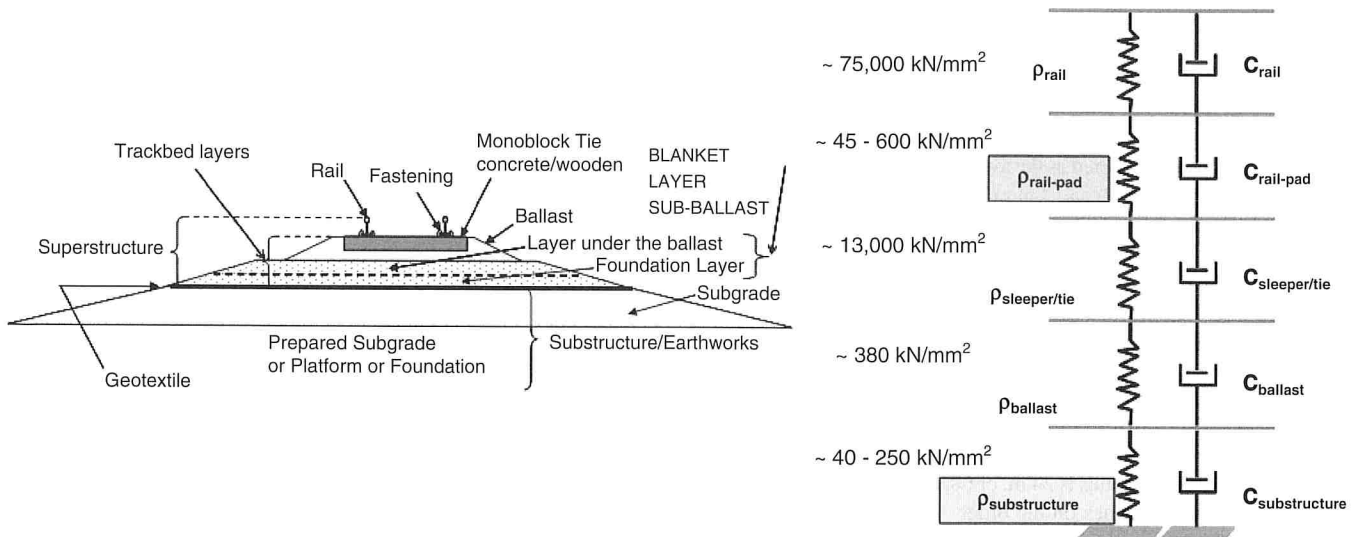


FIGURE 4 Typical cross section of ballasted track and its simulation as combination of springs and dampers and representative values for  $\rho_i$  as derived from measurements.

bearing capacity and the stiffness of the track, according to the following equation [cited also as Equation 3 in Ahlbeck et al. (21) and as Equation 16 in Kerr (19)]:

$$\frac{1}{\rho_{\text{total}}} = \frac{1}{\rho_{\text{rail}}} + \frac{1}{\rho_{\text{pad}}} + \frac{1}{\rho_{\text{sleeper}}} + \frac{1}{\rho_{\text{ballast}}} + \frac{1}{\rho_{\text{substructure}}} \quad (15)$$

The following static stiffness coefficients per layer have been measured on track (Figure 4):

- $\rho_{\text{rail}}$  ranging from 50,000 to 100,000 kN/mm with average of 75,000 kN/mm;
- $\rho_{\text{tie}}$  ranging from 500 to 800 kN/mm for oak wooden tie and 12,000 to 15,000 kN/mm with average of 13,500 kN/mm for a concrete tie;
- $\rho_{\text{ballast}}$  on order of 380 kN/mm for ballast 2 years after laying, relatively “fouled”; and
- $\rho_{\text{substructure}}$  ranging from (a) 20 to 60 kN/mm for pebbly subgrade, (b) 80 to 100 kN/mm in the case of a well-compacted substructure, (c) between 86 and 171 kN/mm for NBS (Neubaustrecke, or new lines’ structure) of the German railways (22), and (d) on the order of 250 kN/mm for the case of a ballast bed of small thickness laid on the concrete base of a tunnel or bridge deck (23).

The tie-pad stiffness  $\rho_{\text{pad}}$  plays a key role in the loading of the tie, and typically its value is estimated from the load–deflection curve provided by the manufacturer, by using a trial-and-error method [for a description of the method, see Giannakos (12)].

As already mentioned, three out of the five layers, namely, the rail, tie, and ballast, contribute only 6% to 10% to the total track stiffness  $\rho_{\text{total}}$ . The total track stiffness is affected mainly by the static stiffness coefficients of the pad,  $\rho_{\text{pad}}$ , and of the substructure,  $\rho_{\text{substructure}}$ . With the new specifications in high-speed lines for the almost undeflected substructure with very stiff behavior, the fastening remains the sole factor controlling the elasticity in the track’s superstructure and thus (as derived from Equation 15) increasing the elasticity of the track system through reduction in the absolute value of the coefficient  $\rho_{\text{total}}$ .

The question arises about whether the high-speed lines with 22.5 t per axle and maximum operational speed  $V \geq 250$  km/h or the heavy haul lines with 35.38 t per axle and 60 mph apply higher action on the track support points.

## COMPARISON OF ACTIONS ON TRACK IN HIGH-SPEED AND HEAVY-HAUL RAILWAYS

### Application of Resilient Fastenings in High-Speed Railroad Track

The following data were used as input for the high-speed lines in the four theoretical methods (AREMA, German, French, Giannakos) for the calculation of actions on the track, derived from the high-speed lines of mixed traffic in Germany (and Greek-designed lines as well): wheel load equals 100 kN (InterCity Express-1, mixed traffic, 112.5 kN for freight trains), maximum operational speed of 250 and 300 km/h, nonsuspended masses equals 1.0 t, height of center of gravity of the vehicle over the rail running table is 1.5 m, superelevation deficiency equals 160 mm, rail type is UIC60 (60 kg/m), and monoblock tie of B70 prestressed concrete (German, used in the Greek network also). For the Nabla fastening the data from the lines in France are used: wheel load equals 85 kN, maximum operational speed is 300 km/h, nonsuspended masses equal 1.0 t, height of center of gravity of the vehicle over the rail running table is 1.0 m, superelevation deficiency equals 160 mm, rail type is UIC60 (60 kg/m), and U41 twin-block concrete tie. The coefficients of stiffness used are  $\rho_{\text{rail}} = 75,000$  kN/mm,  $\rho_{\text{tie}} = 13,500$  kN/mm,  $\rho_{\text{ballast}} = 380$  kN/mm, range of  $\rho_{\text{substructure}} = 40$  to 250 kN/mm (228.41 to 1,427.5 kips/in.) and  $\rho_{\text{fastening}}$  calculated by the load–deflection curves for two kinds of fastenings, both laid in the Greek network, designed for maximum operational speeds of 250 km/h and 300 km/h: W14 with two kinds of pads, Zw700 of Saargummi and Zw700 of Wirtwein, and Nabla with 9-mm pad (laid in TGV lines). The 9-mm pad has been tested in the United States as described in FRA (6). The  $\rho_{\text{fastening}}$  has been calculated separately for each combination of track layers by the

trial-and-error method [see Giannakos (12, 23)]. Maximum operational speeds of 300 km/h were also analyzed because both types of fastenings are laid in European networks with a maximum operational speed of 300 km/h and, consequently, the fastenings undertake the relevant actions on the track. In all cases the situation of the rail running table has been considered as the average of a non-ground rail [see Giannakos and Loizos (14) and Alias (15)]. This means that the most adverse condition of the rail has not been taken into account.

### Stiff Versus Resilient Fastenings in Heavy-Haul Railroad Track

In the United States heavy-haul freight railway transportation has different characteristics: wheel load of 39,006 lb or 17.69 t (35.38 t per axle); maximum speed of 60 mph, that is, 96.6 km/h; and distance between two consecutive ties is 24 in. or 60.96 cm. Ahlbeck et al. (21) and Hay (1) propose values on the order of 4,450 lb or 2,018 t per wheel for the unsprung (nonsuspended) masses. In this paper that is the value used for the unsprung masses. The use of wooden ties equipped with stiff fastenings with no rail pad, such as spikes or similar, on heavy-haul railroad track is compared with use of a very resilient fastening [for a more detailed analysis see Giannakos (23)]. The data taken into account for heavy haul in the United States are wheel load of 39,006 lb or 176.9 kN, maximum operational speed of 60 mph or 96.54 km/h, nonsuspended masses of 2.02 t, height of center of gravity of the vehicle over the rail running table of 1.5 m, superelevation deficiency of 50 mm, rail type of 140RE according to AREMA (10), monoblock tie of prestressed concrete 8 ft 6 in.  $\times$  12 in. and wooden tie 8 ft 6 in.  $\times$  9 in. (10),  $\rho_{\text{rail}} = 75,000$  kN/mm,  $\rho_{\text{tie}} = 13,500$  kN/mm,  $\rho_{\text{ballast}} = 380$  kN/mm, and range of  $\rho_{\text{substructure}} = 40$  to 250 kN/mm. The variable  $\rho_{\text{fastening}}$  is calculated (a) by the load-deflection curves for two kinds of fastenings for concrete ties: W24 with Sk1-24 clip and highly resilient strain-attenuating Zw700WIC rail pads and Safelock and (b) as 7,500 kN/mm for steel base plate with spikes in wooden ties. The  $\rho_{\text{fastening}}$  for concrete ties has been calculated separately for each combination of track layers by the trial-and-error method [see Giannakos (12, 23)]. The aforementioned average values of  $\rho_{\text{rail}}$ ,  $\rho_{\text{tie}}$ , and  $\rho_{\text{ballast}}$  and the variation of  $\rho_{\text{substructure}}$  have been measured on the track by the German railways and have been accepted and used by the French railways as well (17). From these measurements the static stiffness coefficient for the wooden tie is 800 kN/mm and the static stiffness coefficient of the steel base plate is derived as equal to the static stiffness coefficient of the steel tie, 7,500 kN/mm.

In both cases, for high-speed and heavy-haul railways, to calculate the acting forces on the superstructure and the ties by applying the aforementioned equations in a multilayered construction with poly-parametrical function, the exact rigidity of the elastic pad of the fastening for each combination of parameters must be determined. In the case of the resilient fastenings, the pad's most adverse load-deflection curve is considered because it describes the behavior of the pad during the approach of the wheel. The stiffness of the substructure varies from 40 kN/mm for a pebbly substructure to 250 kN/mm for a rocky tunnel bottom or concrete bridge with insufficient ballast thickness. Each time this stiffness changes in Equation 15 above, the acting stiffness of the tie pad also changes.

### Results

Figures 5 through 8 depict the results of the four aforementioned methods. According to these figures the actions and reactions on

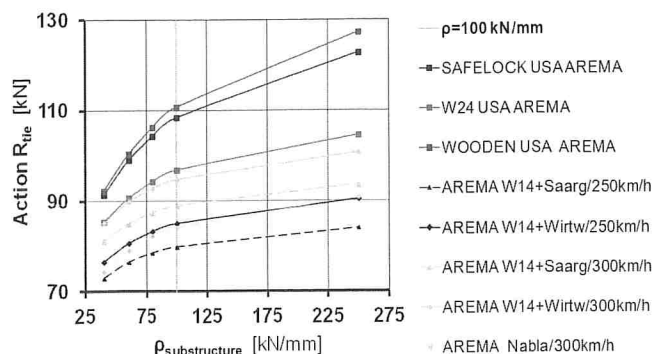


FIGURE 5 Actions on each track's support point (tie) according to method cited in AREMA (Saarg = Saargummi; Wirtw = Wirtwein).

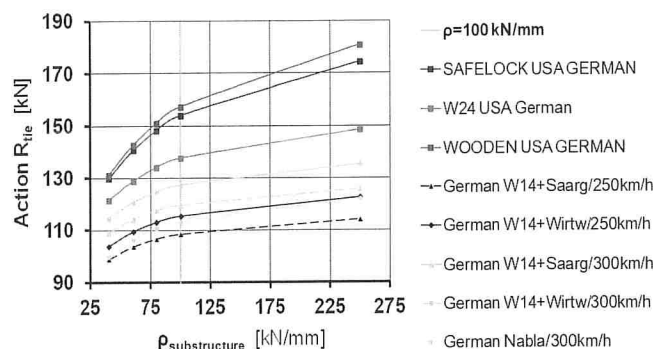


FIGURE 6 Actions on each track's support point (tie) according to method cited in German literature.

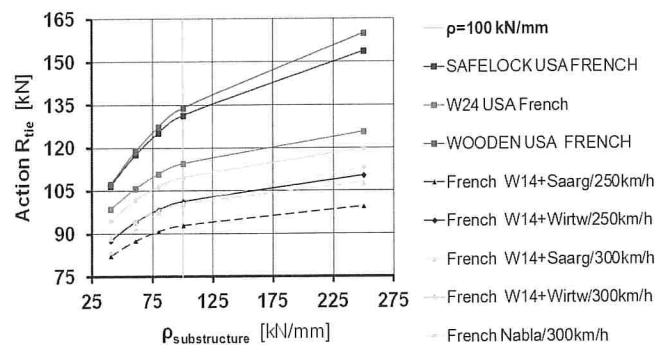


FIGURE 7 Actions on each track's support point (tie) according to method cited in French literature.

each tie for heavy-haul and high-speed lines are much smaller in the case of resilient fastenings and concrete ties compared with the stiff fastenings and wooden ties [see relevant analysis in Giannakos (23)]. An examination of the results for the resilient fastenings from each method yields the following information.

#### AREMA Method

The Safelock pad, which is of medium stiffness and elasticity, generates higher reactions per tie than does the W24, ranging from 6.9% for pebbly subgrade ( $\rho_{\text{substr}} = 40$  kN/mm and  $\rho_{\text{pad}} = 400$  kN/mm for Safelock and  $\rho_{\text{pad}} = 85.2$  kN/mm for W24) to 17.4% for rocky tunnel bottom or concrete bridge with insufficient ballast

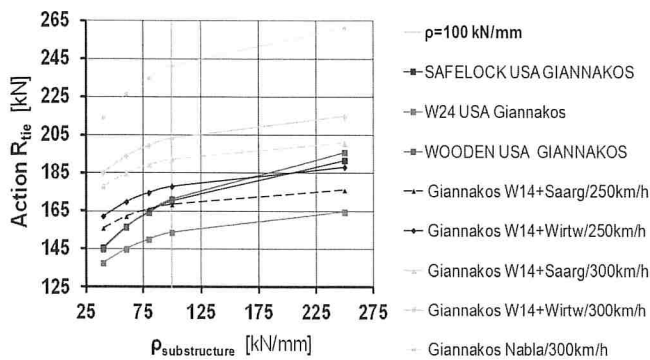


FIGURE 8 Actions on each track's support point (tie) according to Giannakos (2004) method.

depth ( $p_{\text{substr}} = 250 \text{ kN/mm}$  and  $p_{\text{pad}} = 389 \text{ kN/mm}$  for Safelock and  $p_{\text{pad}} = 91.5 \text{ kN/mm}$  for W24). The decisive value of the subgrade is between  $40 \text{ kN/mm}$  and  $100 \text{ kN/mm}$ . In general the very resilient W24 fastening for heavy-haul lines generates actions per tie close to the values of the three examined fastenings at the high-speed lines of  $300 \text{ km/h}$  (from  $+5.3\%$  to  $+8.9\%$ ). Safelock generates higher actions compared with Nabla (wheel load  $85 \text{ kN}$ ), ranging from  $+22.85\%$  for a pebbly substructure ( $p_{\text{substr}} = 40 \text{ kN/mm}$ ) to  $+28.70\%$  for an excellent subgrade of  $p_{\text{substr}} = 100 \text{ kN/mm}$ . Safelock, when compared with the W14 + Saargummi pad, generates higher actions, ranging from  $+12.53\%$  for pebbly substructure ( $p_{\text{substr}} = 40 \text{ kN/mm}$ ) to  $+21.82\%$  for an excellent subgrade of  $p_{\text{substr}} = 100 \text{ kN/mm}$ .

#### German Method

The Safelock pad generates reactions that are almost identical to the actions generated by spikes and wooden ties (difference from  $-1\%$  to  $-3.72\%$ ). In general the very resilient fastening W24 for heavy haul generates actions per tie close to the values of the three examined fastenings on the high-speed lines of  $300 \text{ km/h}$  (from  $+11.3\%$  to  $+15.1\%$ ).

#### French Method

Results of the French method (covering a probability of occurrence of  $95.5\%$ ) are similar to the German method results (for  $99.7\%$  probability), but with values slightly lower than those calculated with the German method.

#### Giannakos (2004) Method

The actions on the track for high-speed lines of  $300 \text{ km/h}$  and with a  $100\text{-kN}$  wheel load are significantly higher than the actions in heavy haul, ranging from  $+22.2\%$  (W14 + Saargummi) compared with Safelock and  $29.44\%$  compared with W24, for pebbly substructure ( $p_{\text{substr}} = 40 \text{ kN/mm}$ ), to  $+12.9\%$  (W14 + Saargummi) compared with Safelock and  $25\%$  compared with W24, for excellent subgrade of  $p_{\text{substr}} = 100 \text{ kN/mm}$ . This method as cited below is closer to the real track conditions.

The Giannakos method was developed as a result of an almost 10-year research program under the author's guidance in the Greek railway network (with the participation of the French railways, a subsidiary of the Belgian railways, and universities in Greece,

Austria, etc.) to investigate the causes of the appearance of extensive cracking in concrete ties of French technology (more than  $60\%$  of the total number laid on the track). Notably, the laboratory tests showed, beyond any doubt, that the cracked ties in the Greek network were produced in full compliance with the existing prescriptions and technical specifications of the time. Moreover the tie samples—chosen randomly from the track—presented strength values in laboratory tests higher than prescribed in the specifications. The cracking was not a result of defective manufacture of the original ties or of noncompliance with the specifications. The existing international bibliography (American, German, and French) includes various methods that suggest respective formulas for a realistic estimation of the actions on the ties.

The load that is derived when these formulas (AREMA, German, French) are applied under the most adverse conditions gives values that justify none at all or sporadic appearance of cracks (on the order of  $1\%$  to  $2\%$ ) but do not justify at all their systematic appearance at  $60\%$  of the ties (and more) in the Greek railway network in the 1980s. The Giannakos method foresees the extended cracking as it was observed in practice (at a percentage  $>60\%$ ). The typical crack appeared at the middle of the concrete block, with direction from the lower tie's seating surface (on the ballast) toward the upper surface of the tie at the position of the rail seat (see Figures 1 and 2). The cracks were generated as a result of factors of railway operation that aggravate the situation of the track (actions on track panel), underestimation of the real actions on ties, inadequate reinforcement, and use of nonappropriate (smooth) steel bars. These formulas, however, do not justify such an extensive appearance of cracks. By applying the aforementioned methodologies to calculate the actions on sleepers, the results as depicted in Figure 9 are derived [for a detailed description see Giannakos and Loizos (14)]. Thus the Giannakos method is closer to real conditions for calculating actions on the track.

In Germany an investigation was performed that led to results in a similar direction as far as the different factors influencing the actions on the track were concerned [see Müller-Borutta et al., 25]. The International Federation of Concrete invited the author to participate in its Task Group 6.5 to draft the publication *Precast Concrete Railway Track Systems*, which covers concrete ties extensively (prefabricated) (26). In this publication only the Giannakos (2004) method is

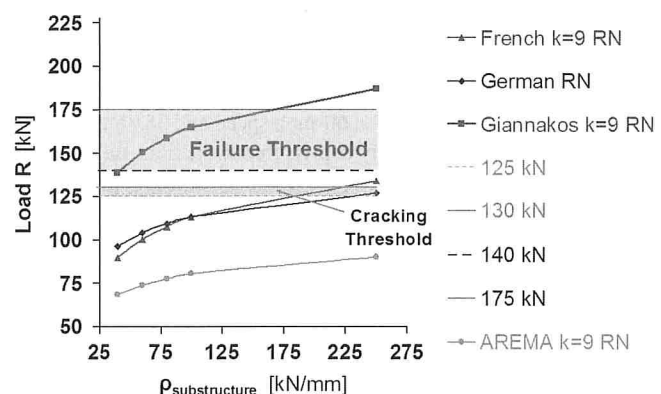


FIGURE 9 Calculation of actions on U2 and U3 twin-block ties with four methods: (a) method cited in French literature [Alias (15)], (b) method cited in German literature [Eisenmann (24)], (c) method cited in AREMA (10) and Hay (1), and (d) Giannakos method (12). Only Giannakos method predicts extended cracking and failure of U2 and U3 ties (more than  $60\%$  of total number on the track).

TABLE 1 Total Acting Loads on Track  $Q_{total}$  According to Four Methods

			Giannakos and French Methods					
Type of Tie + Fastening	AREMA Method (kN/mm)	German Method (kN/mm)		Subgrade, $\rho_{\text{subgr}}$ (kN/mm)				
				40	60	80	100	250
Heavy-haul wooden + spikes	266	378	Giannakos	249	255	259	263	276
			French	229	233	236	238	248
Heavy-haul concrete + safelock	266	378	Giannakos	250	256	260	263	263
			French	230	234	237	239	247
Heavy-haul concrete + W24	266	378	Giannakos	246	250	252	254	260
			French	227	230	232	233	237
High-speed line + W14 + Wirtwein 250 km/h	230	312	Giannakos	232	237	240	243	250
			French	195	199	201	203	208
High-speed line + W14 + Saargummi 250 km/h	230	312	Giannakos	238	243	247	249	256
			French	193	195	197	198	202
High-speed line + Nabla 300 km/h <sup>a</sup>	218	293	Giannakos	223	231	237	240	252
			French	181	187	191	193	201
High-speed line + W14 + Wirtwein 300 km/h	256	344	Giannakos	289	297	265	268	277
			French	211	215	218	220	226
High-speed line + W14 + Saargummi 300 km/h	256	344	Giannakos	249	254	258	260	266
			French	216	220	223	224	230

<sup>a</sup>Static wheel load ( $Q_{wheel}$ ) = 85 kN; nonsuspended masses (NSM) = 1t.

cited and referenced for the estimation of acting loads in the precast concrete railway track systems.

This paper also proposes that the results of the Giannakos (2004) method are closer to real track conditions and should be considered to be more reliable. In any case an investigation with measurements on the track should be performed for the heavy-haul railroads. In this method (a) the statistical certainty covered is 99.7% taking into account three times the standard deviation of the (random) dynamic component of the load and (b) the coefficient of distribution for the dynamic component of the total load is not distributed to the adjacent ties because the static and semistatic (due to the superelevation deficiency) components are distributed through the coefficient  $A_{dynam}$  instead of the  $A_{stat}$ . These assumptions are based on the fact that because of its eigen frequency, the track response is not sensitive to the frequency of the running axle and, therefore, the total dynamic component acts on one tie.

From an evaluation of the results of the four methods, one should conclude that the actions on the track are at least of comparable magnitude for both cases: high-speed lines with an axle load of 22.5 t and heavy-haul railroads with a 60-mph speed and 35.38 t per axle. The Giannakos method clearly depicts that the actions on the track due to heavy haul are significantly less than in high-speed lines. The aforementioned results lead to the conclusion that the tests (included in the railway regulations) for the approval of resilient pads and clips should be of a similar order for both cases or that the correctly designed and produced pads and clips for high-speed lines should be proper for heavy-haul railways also. The life cycle of the pads and fastening clips is dependent on the actions exerted on each tie per passing axle and the number of axle passes. Because the actions are of comparable magnitude, the tests for high speed are enough for heavy haul also. Harrison and Ahlbeck verified the much better behavior on the track and in the lab of the resilient versus the stiff pads (10). According to the Greek and German regulations, the pads should provide a life cycle of 20 years (or 400 million tonnes) (DB-TL 918 235).

The corresponding very resilient fastening clip should provide a life cycle of 30 years (or an estimation of 600 million tonnes). Moreover, according to research performed by the European Rail Research Institute of the International Union of Railways (ERRI-U.I.C.), the W clip (e.g., the W14 or W24 clip) stops at a 2-mm rail tilt deflection because of its shape and design and presents highly sufficient strength [see ERRI (27, 28)]. The development of new, highly resilient pads permitted, technologically, an increase in the pad's elasticity on the order of 45 to 50 kN/mm (256.96 to 285.51 kips/in.) because of the air circulation that facilitates the heat removal and the (probable) water sewerage (23). This should imply an avoidance of the pad degrading and blowing out as well as the rail-seat abrasion.

Moreover, in the case in which the total acting load on track  $Q_{total}$  is calculated, the results are of a similar magnitude for all four methods for heavy-haul and high-speed lines, as depicted in Table 1. In this case the load is distributed along the track and generates the maximum  $R$  per tie.

## CONCLUSION

The development of modern, highly resilient fastenings significantly reduces actions on the concrete ties and the track superstructure. There is a need for the resilient fastenings to be used in the modern railway tracks because they contribute elasticity in the track performance and counterbalance the rigidity of the modern, almost undeflected, railway substructure. The parametric investigation of different kinds of resilient fastenings in heavy-haul and high-speed lines for fluctuation of the substructure from a very elastic to a very rigid subgrade showed that the actions on the track in both cases are of comparable magnitude for the methods cited in the American, French, and German literature. According to the Giannakos (2004) method the actions in high-speed lines are significantly higher than in heavy haul. All these results lead to the conclusion that the tests (included in the railway regulations) for the approval of resilient pads and clips should be of a

similar order for the two cases. Moreover, the properly designed pads and clips produced for high-speed lines should also be effective in the case of heavy-haul railways. Highly resilient pads of such characteristics have been produced and used in high-speed lines under exploitation. They are “channeled” or “studded.” The development of channeled or studded pads permitted, technologically, an increase in the elasticity of the pads (production of very resilient pads) on the order of 45 to 50 kN/mm (256.96 to 285.51 kips/in.) because of the air circulation that facilitates the heat removal and the (probable) water sewerage (12). According to the Greek and German regulations [DB-TL 918 235] the pads should provide a life cycle of 20 years or 400 million tonnes. The corresponding highly resilient fastening clip should provide a life cycle of 30 years (23). This is important because the life cycle of the pads and the fastening clips is dependent on the actions exerted on each tie per passing axle and the number of axle passes.

## REFERENCES

- Hay, W. *Railroad Engineering*. John Wiley & Sons, New York, 1982.
- Giannakos, K. *The Use of Strain Attenuating Tie Pads and Its Influence on the Rail Seat Load in Heavy-Haul Railroads*. JRC-2010, Urbana-Champaign, Ill., April 27–29, 2010.
- Optimum Adaptation of the Conventional Track to Future Traffic*. D117. Rapports 1–27. Office for Research and experiments (ORE), International Union of Railways (UIC), Utrecht, Netherlands.
- Eisenmann, J. *Schotteroberbau—Möglichkeiten und Perspektiven für die Moderne Bahn*, 1988.
- Winkler, E. *Die Lehre von der Elastizität und Festigkeit (The Theory of Elasticity and Stiffness)*. H. Dominicus, Prague, Bohemia, 1867.
- Investigation of the Effect of Tie Pad Stiffness on the Impact Loading of Concrete Ties in the Northeast Corridor*. Final report. Office of Research and Development, FRA, U.S. Department of Transportation, April 1983.
- Eisenmann, J., and L. Mattner. *Auswirkung der Oberbaukonstruktion auf die Schotter—und Untergrundbeanspruchung*. *Eisenbahningenieur*, Vol. 35, No. 3, 1984.
- Giannakos, K. *Loads on Track, Ballast Fouling, and Life-Cycle under Dynamic Loading in Railways*. *International Journal of Transportation Engineering—ASCE*, Vol. 136, No. 12, 2010, pp. 1075–1084.
- Zimmermann, H. *Die Berechnung des Eisenbahnoberbaues*. Verlag von Wilhelm Ernst & Sohn, Berlin, 1941.
- Manual for Railway Engineering*, rev. ed. American Railway Engineering and Maintenance-of-Way Association, Lanham, Md., 2005.
- Selig, E., and J. Waters. *Track Geotechnology and Substructure Management*. Thomas Telford, London, 1994 (reprinted 2000).
- Giannakos, K. *Actions on the Railway Track*. Papazisis, Athens, Greece, 2004. [www.papazisi.gr](http://www.papazisi.gr).
- Harrison, H., and D. Ahlbeck. *Railroad Track Structure Performance under Wheel Impact Loading*. In *Transportation Research Record 1131*, TRB, National Research Council, Washington, D.C., 1987, pp. 81–88.
- Giannakos, K., and A. Loizos. *Evaluation of Actions on Concrete Sleepers as Design Loads—Influence of Fastenings*. *International Journal of Pavement Engineering*, Nov. 2009.
- Alias, J. *La voie ferrée*, Ileme ed., Eyrolles, Paris, 1984.
- Prud'homme, A., and J. Eriau. *Les Nouvelles Traverses en beton de la S.N.C.F. Revue Generale des Chemins de Fer*, 1976.
- Prud'homme, A. *Sollicitations statiques et dynamiques de la voie*. In *Direction des Installations Fixes*, S.N.C.F., 1966.
- Prud'homme, A. *La Voie. Revue Generale des Chemins de Fer*, 1970.
- Kerr, A. *The Determination of the Track Modulus k for the Standard Track Analysis*. *Proc., AREMA 2002 Annual Conference*.
- Bathurst, L., and A. Kerr. *An Improved Analysis for the Determination of Required Ballast Depth*. *Proc., AREMA 1999 Annual Conference*.
- Ahlbeck, D. R., H. C. Meacham, and R. H. Prause. *The Development of Analytical Models for Railroad Track Dynamics*. In *Proc., Railroad Track Mechanics and Technology Symposium Held at Princeton University, April 21–23, 1975* (A. D. Kerr, ed.), pp. 239–264.
- Eisenmann, J., G. Leykauf, and L. Mattner. *Vorschläge zur Erhöhung der Oberbauelastizität*. *ETR*, Vol. 43, No. 7/8, 1994.
- Giannakos, K. *Heavy Haul Railway Track Maintenance and Use of Resilient Versus Stiff Fastenings*. Presented at 90th Annual Meeting of the Transportation Research Board, Washington, D.C., 2011.
- Eisenmann, J. *The Rail as Support and Roadway, Theoretical Principles and Practical Examples*. In *Railroad Track—Theory and Practice* (F. Fastenrath, ed.), Frederic Ungar, New York, 1981.
- Müller-Borutta, F. H., D. Ebersbach, and N. Breitsamter. *Dynamische Fahrplanmodelle für HGV-Strecken und Folgerungen für Komponenten*. *Eisenbahntechnische Rundschau*, Vol. 47, No. 11, Nov. 1998.
- Precast Concrete Railway Track Systems*, spec. ed., International Federation of Concrete, Task Group 6.5 (invited member K. Giannakos) of Committee 6 for Prefabrication, federation internationale du beton, Switzerland, 2006.
- Calculation of Twist and Displacements in UIC60 Rail Using Rail Fastening System W and Rail Pads Zw687a and Zw700*. D 170/DT 282. ERRI, Utrecht, Netherlands, Sept. 1994.
- Stresses on the Fastening System under the Action of Wheel Loads*. D 170/DT 302. ERRI, Utrecht, Netherlands, Sept. 1994.

*The Railroad Track Structure System Design Committee peer-reviewed this paper.*



# Movement of Water Through Ballast and Subballast for Dual-Line Railway Track

Gurmel S. Ghataora and Ken Rushton

The purpose of effective track design is to ensure that load from trains can be safely supported by subgrade soils, which are likely to be affected by water. Because it is almost impossible to prevent water from entering the ballasted track and therefore affecting the underlying layers, it is vitally important that water can drain away rapidly. The study undertaken shows how water moves through a dual-railway track, which is either symmetrical about the centerline or superelevated with a continuous slope of the subgrade across both tracks. The study shows that water may be retained in the track nearest the drain for more than a week with possible consequential impact on increased deformation and therefore the need for more maintenance. It also shows that a high permeability composite may be included near the base of the subballast to effect significant improvement in track drainage.

Selig and Waters describe various geotechnical aspects of railway track design and maintenance (1). They explain the functions of each track layer (ballast, subballast, and subgrade) and note the important requirement for adequate drainage of both the ballast and subballast layers following rainfall. The presence of fines in the ballast—arising from any or all of the following: ballast breakdown through ballast-tamping and attrition (perhaps the largest source), wear of the sleepers, and migration of the fines from the subgrade soils—leads to reduction in permeability of the ballast. Hyslip and McCarthy note that one of the critical factors that lead to railway substructure problems is “inadequate drainage of the track substructure allowing water to remain in contact with the clay subgrade for extended periods of time” (2). Thus, it is important to ensure that water is removed from the top of the subgrade in a reasonable time. The effect of water on the strength of soils depends on the type of clay because some soil types are more susceptible to softening than are others. In the presence of water and under repeated loading, the uppermost thin upper layer of the subgrade is affected most. Depending on the clay type, even stiff subgrade soils can soften in this environment [Ghataora et al. (3)].

Movement of water through the railway track, with respect to time, was investigated experimentally by Heyns in 2000 (4). His study showed the process of the flow of water, arising from rainfall, in ballast and subballast. It was shown that water is shed rapidly through the ballast layer and then more slowly through the subballast when rainfall ceases. He also showed that water was shed

more slowly through horizontal subgrade compared with instances in which the subgrade is sloping. On the basis of the Dupuit-Forchheimer approximation, a time variant numerical model was developed by Youngs and Rushton (5, 6) to simulate the observations made by Heyns (4). The differential equation, Equation 1, describes the flow through the ballast and sand blanket to the downstream drain with three components. The left-hand expression describes the flow through the sand blanket and the ballast; the first term on the right-hand side defines water taken into or released from storage as the water table rises or falls. The final term describes any recharge entering the system

$$\frac{\partial}{\partial x} [K_s t_s + K_b (h - z_{sg} - t_s)] \frac{\partial h}{\partial x} = S_y \frac{\partial h}{\partial t} - q \quad (1)$$

where

$h$  = water table elevation,  
 $K$  = permeability,  
 $s$  and  $b$  = sand blanket and ballast,  
 $t_s$  = thickness of sand blanket,  
 $z_{sg}$  = elevation of subgrade above lowest point of subgrade,  
 $q$  = recharge,  
 $S_y$  = specific yield (drainable porosity),  
 $t$  = time, and  
 $x$  = horizontal coordinate.

More details of the numerical model are given by Youngs and Rushton (5, 6).

The model was able to represent flow in the ballast and subballast to simulate water levels in the case of both a horizontal and a sloping subgrade [see Rushton and Ghataora (7)]. For the sloping ballast the generalized track model was based on the subballast dipping toward a ditch in the cess as shown in Figure 1. In Figure 1, diagrams  $b$ ,  $c$ , and  $d$ , comparisons are made between Heyns' experimental results and the results from the numerical model; the satisfactory agreement demonstrates that the model can represent the outflows and the water table elevations. Numerical simulations were used to examine the movement of water in a typical railway track section shown in Figure 2.

Ballast fouling was examined, and it was found that with moderately clean ballast, as described by Selig and Waters, permeability reduces by up to 90% and the fall in the water level is much slower, as can be seen in Figure 3 (1). Similarly, the effect of a sloping subgrade on the water table is shown in Figure 4.

Following this study a further investigation was undertaken by Rushton and Ghataora to examine the effectiveness of a highly permeable layer in the subballast (8). They concluded that the best position for the highly permeable layer was at the base of the sub-

School of Civil Engineering, University of Birmingham, Edgbaston, Birmingham B15 2TT, United Kingdom. Corresponding author: G. S. Ghataora, g.s.ghataora@bham.ac.uk.

*Transportation Research Record: Journal of the Transportation Research Board*, No. 2289, Transportation Research Board of the National Academies, Washington, D.C., 2012, pp. 78–86.  
 DOI: 10.3141/2289-11

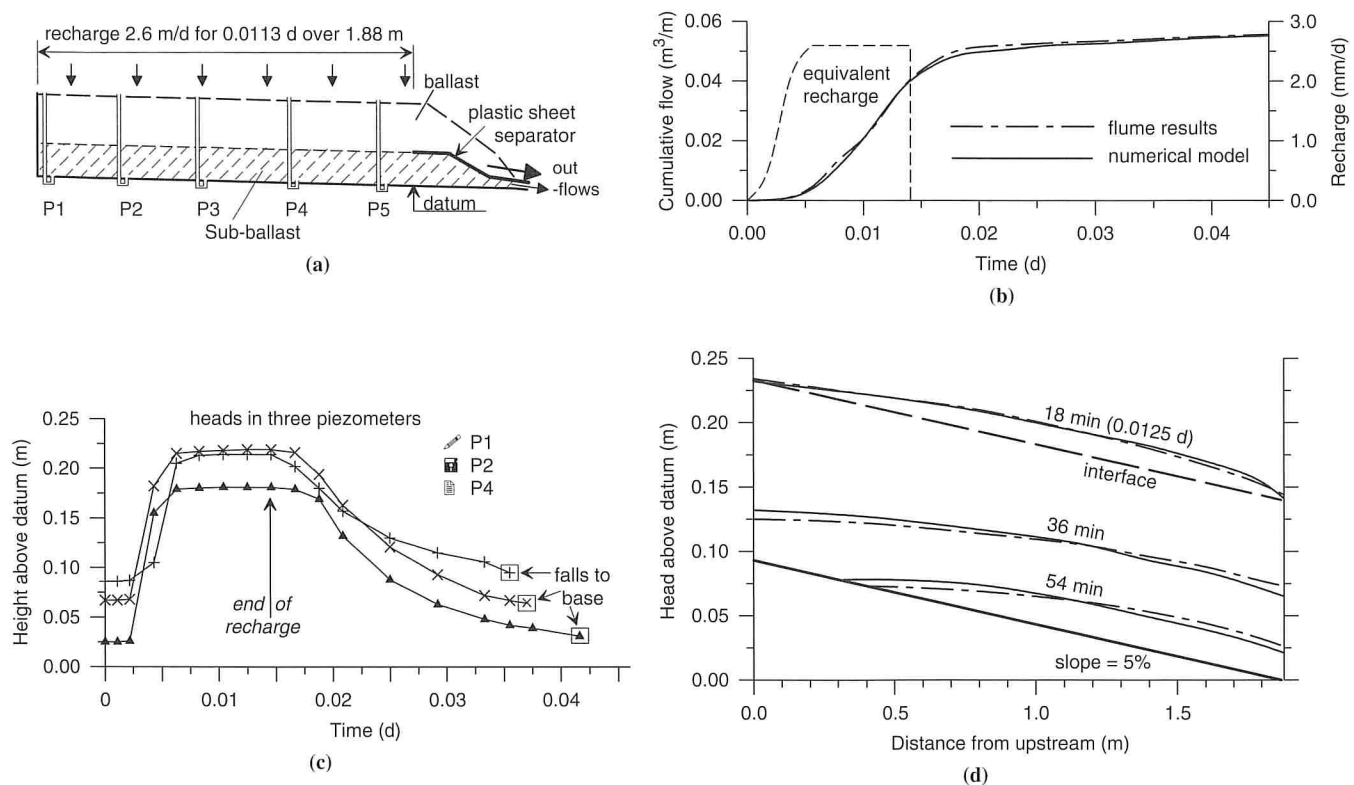


FIGURE 1 Apparatus used by Heyns and experimental and numerical results for slope of 5% with rainfall of 2.6 m/day (m/d) for 16 min: (a) sketch of flume test, (b) change in piezometric head with full line indicating numerical results, (c) cumulative outfall, and (d) selected piezometric results on cross section. [Source: Rushton and Ghataora (7).]

ballast. This conclusion is demonstrated in Figure 5, which shows the relationship of the time required for the water level to drop to 1 cm above the subgrade and the ratio ( $K/S_y$ ) of permeability to specific yield (drainable porosity) for high-permeability composite at the base of the subgrade and without a geocomposite.

Permeability is a measure of the rate at which water moves through the soil, and specific yield ( $S_y$ ) is the volume of water that can drain due to a unit fall in head per unit plan area. Thus, a larger specific yield means a greater volume of drainable water. Clays may have a specific yield as low as 0.02, whereas coarse sand and gravels may have an  $S_y$  of up to 0.20. In the examples shown in Figures 3, 4, and 5, ballast is assumed to have 10 times more drainable water compared with subballast. The ratio of  $K/S_y$  indicates the velocity of flow and

drainable quantity. So for a given material, a higher value of the  $K/S_y$  ratio means more rapid drainage.

For geocomposite at the top of the subballast, results are similar to results without the composite. This similarity is not surprising because the water level drops rapidly through the ballast, below the geocomposite, and then slowly through the subballast. Thus, a high-permeability composite above the subballast does little to improve the situation. However, the composite proves to be very effective when it is placed directly above the subgrade layer.

A design chart showing a comparison of the time required to drain water to reference value, without any drainage measures and with permeable geocomposite located above the subgrade, is shown in Figure 5. Results clearly show that including the drainage composite

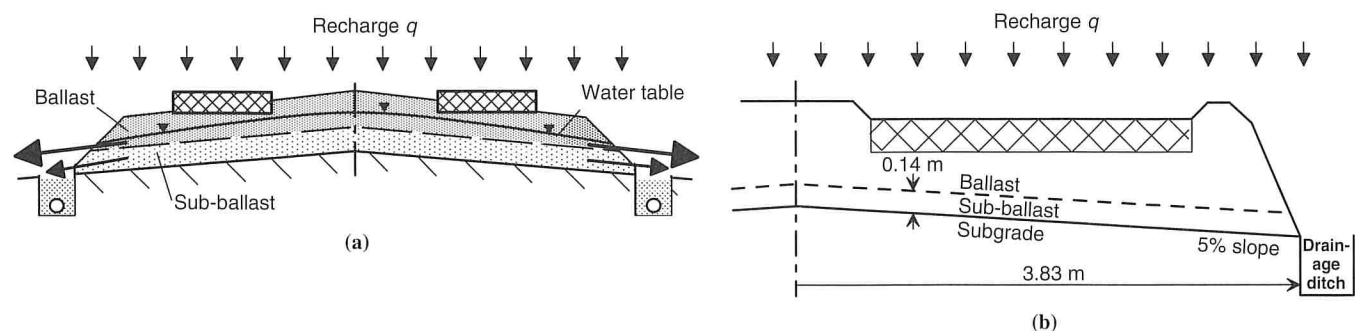


FIGURE 2 Generalized cross sections (7).

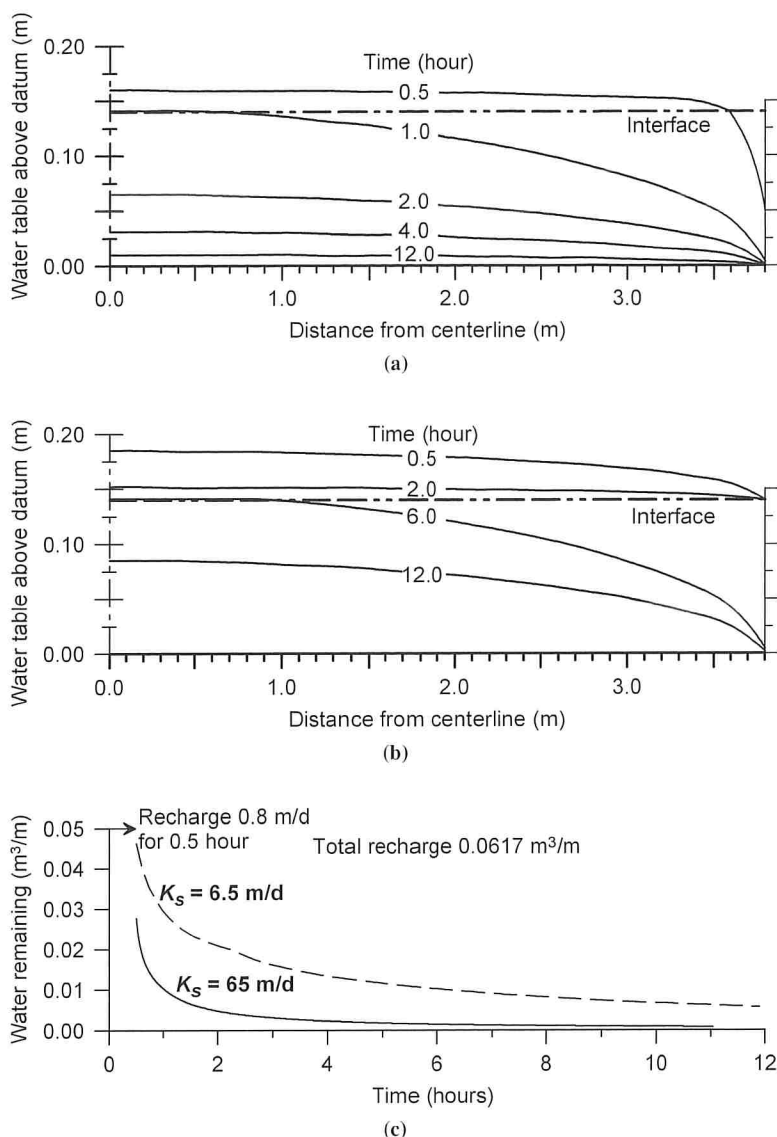


FIGURE 3 Effect of reduced permeability due to fouling (horizontal formation: recharge 0.8 m/day for 30 min; subballast permeability,  $K_s = 65$  m/day or 6.5 m/day and  $S_y = 0.025$ ; ballast permeability,  $K_b = 500 \times K_s$ , specific yield,  $S_y = 0.25$ ): (a) water table elevation, clean subballast,  $K_s = 65$  m/day; (b) water table elevations for moderately clean subballast,  $K_s = 6.5$  m/day; and (c) water remaining in ballast and subballast (7).

above the subgrade significantly reduces the time required for water levels to drop to the reference value (1 cm above subgrade) for  $K/S_y$  ratios of up to about 300.

Rushton and Ghataora's results are for typical levels of British rainfall (7, 8).

## REPRESENTATION OF DUAL-TRACK PROBLEM

Following the recent studies by Rushton and Ghataora a further investigation was undertaken to assess the movement of water through a dual-track arrangement, in which water from the track is assumed to flow to a drain in the cess for the lowest track as shown in Figure 6 (7, 8). This is based on a dual section of track shown in the *AREMA*

*Manual of Railway Engineering* (9). This paper describes the findings of this study.

Precipitation recharge is based on rainfall for the South in the United States as this probably represents the worst-case scenario. Rainfall intensity for 1 h with a 300-year return period is 5.0 in. (10). This intensity is equivalent to 3.05 m/day for 0.0417 d (127 mm in 1 h). In the central-west United States the intensity for 1 h is 3.75 in. For 100-year return periods the intensity is lower.

Initially, symmetrical tracks are considered (Figure 2) with subgrade slopes of 1:24, subballast thickness of 0.23 m (9 in.), and a ballast thickness of 0.23 m (9 in.). Each section has a total width of 4.0 m (13 ft 1.5 in.). For the subballast, the permeability is 1.73 m/day ( $2.0 \times 10^{-5}$  m/s) and  $S_y = 0.05$ ; for the ballast, the permeability is 1,000 m/day ( $1,157 \times 10^{-5}$  m/s) and  $S_y = 0.20$ .

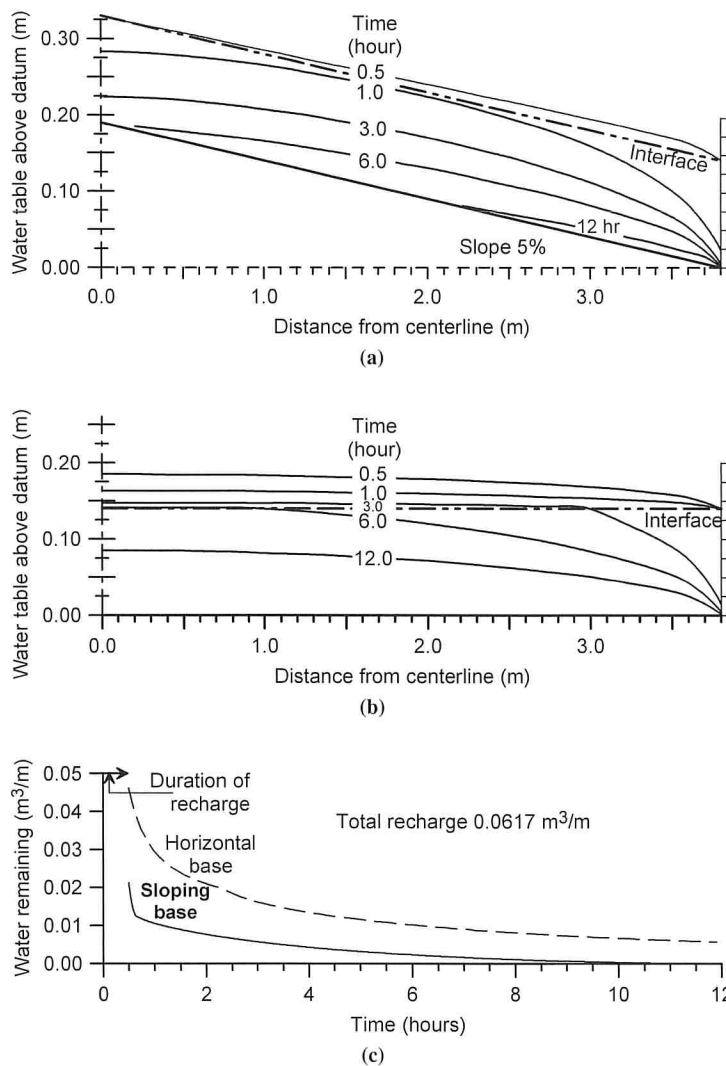


FIGURE 4 Comparison of results for sloping and horizontal subgrade surface (recharge 0.8 m/day for 30 min; subballast permeability,  $K_s = 6.5$  m/day and  $S_y = 0.025$ ; ballast permeability,  $K_b = 500 \times K_s$ , specific yield,  $S_y = 0.25$ ): (a) water table elevation on cross section, 5% sloping base; (b) water table elevations on cross section, horizontal base; and (c) water remaining in ballast and subballast (7).

This was followed by a study of superelevated dual track with ballast beneath each track as shown in Figure 6 and in the cross-section drawing in Figure 7. As for the single track, the subgrade slope is 1:24; the subballast thickness is 0.23 m (9 in.) with ballast thickness of 0.23 m. The total width of the two tracks considered is 8 m, as shown in Figure 7. There is a trench drain in the cess adjacent to the lower track, but water can also drain from the upper shoulder and possibly from between the tracks.

## RESULTS AND DISCUSSION

### Symmetrical Double Track

Results of the movement of water in one half of the track structure for symmetrical dual track are shown in Figure 8. After the end of recharge, that is, the precipitation period, water in the ballast layer

drains away rapidly. Water in the subballast layer drains more slowly and takes about 4.4 days to reduce to 1 cm depth above the subgrade at the midpoint of each track (1 cm depth of water is chosen as an arbitrary reference value). During this period, in particular in the first few days, the presence of water in the subballast layer may have a significant impact on track stiffness and deformation for a road subbase as shown by Hornych et al. (see Figure 9) (11). High track deformations can accelerate ballast deterioration through attrition and result in increased maintenance costs [see Hunt (12)].

There are a number of factors that lead to a slower fall in the water table compared with the initial studies described by Rushton and Ghataora (7). The slope of the single track examined here is 1:24, compared with a slope of 1:20 for the initial studies. Furthermore the permeability of the subballast is substantially lower than for the initial studies. The increased thickness of the subballast, 0.23 m for the current study compared with 0.14 m for the initial investigation, also leads to longer times for the fall in the water table. This is

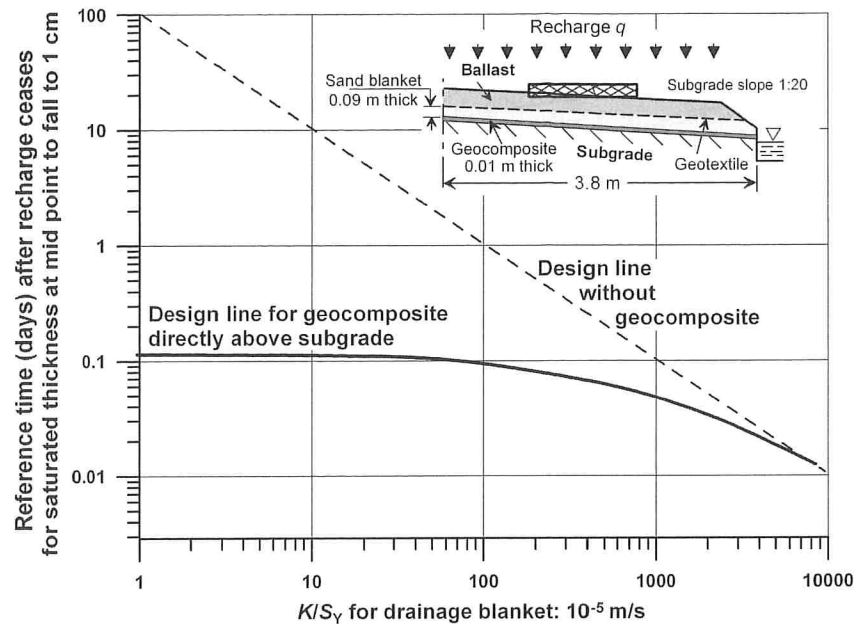


FIGURE 5 Design chart indicating reference time, in which water table falls to 1 cm above subgrade at midpoint of one track, as a function of  $K/S_y$  for subgrade slope of 5%, recharge 0.8 m/day for 0.5 h (0.0208 day); sand blanket, 0.1 m thick,  $K$  and  $S_y$  vary; ballast  $K = 10,000$  m/day (0.116 m/s),  $S_y = 0.20$ . Full line geocomposite located directly above subgrade, 1 cm thick,  $K = 300 \times 10^{-5}$  m/s (259 m/d),  $S_y = 0.05$ . Broken line is for no geocomposite (8).

compounded by the higher rainfall intensity for the current study. A sensitivity analysis described in the second paper by Rushton and Ghataora highlights the manner in which different parameters influence the time for the water table to drain from the subballast (8).

For the superelevated dual-track problem (see the inset of Figure 7) water can discharge from the ballast at the left-hand side, the right-hand side, and on either side of the "trough" in the middle where no ballast is present. Figure 7a illustrates the rate at which the water table falls at the midpoint of the lower track; the full line refers to the situation in which water ponds between the tracks, the broken line to the situation in which water can drain from between the tracks. If water is able to drain away through the trough between the tracks, there will be a lesser quantity to flow through the subballast; consequently drainage is substantially more rapid.

Figure 7b shows detailed results for the situation in which water ponds between the tracks. For the left-hand and right-hand tracks the head of water rises into the ballast during recharge. Soon after recharge

ceases the water table falls and is just above the interface between the ballast and subballast. Thereafter the water table falls slowly in the subballast; the water table reaches the top of the subgrade at the left-hand side after about 6 days, but at 10 days there is still water above the subgrade for more than half of the cross section with the water table at the midpoint on the right-hand track 0.035 m above the subgrade (see Figure 7a).

### Design Curve

Figure 10 includes a design curve showing the effect of the permeability of the subballast on the reference time for the water table to fall to within 1 cm of the subgrade at the midpoint of one track of a symmetrical track arrangement. This diagram refers to specimen American conditions and is based on two values of the specific yield. Although specific yield values are estimations, they do show that as

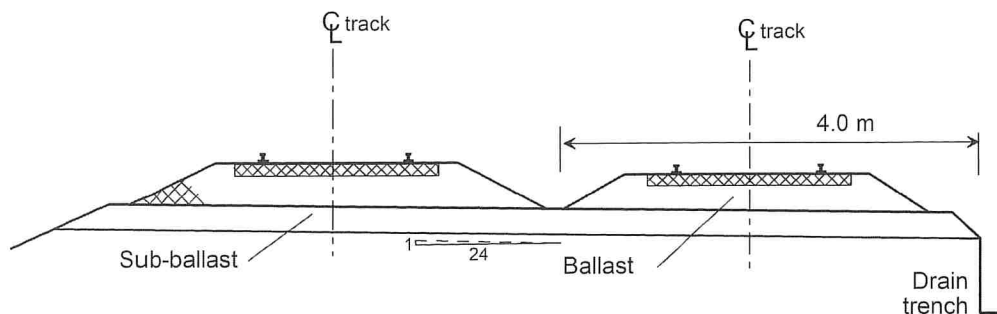


FIGURE 6 Generalized dual-track layout. [Source: Adapted from AREMA (9).]



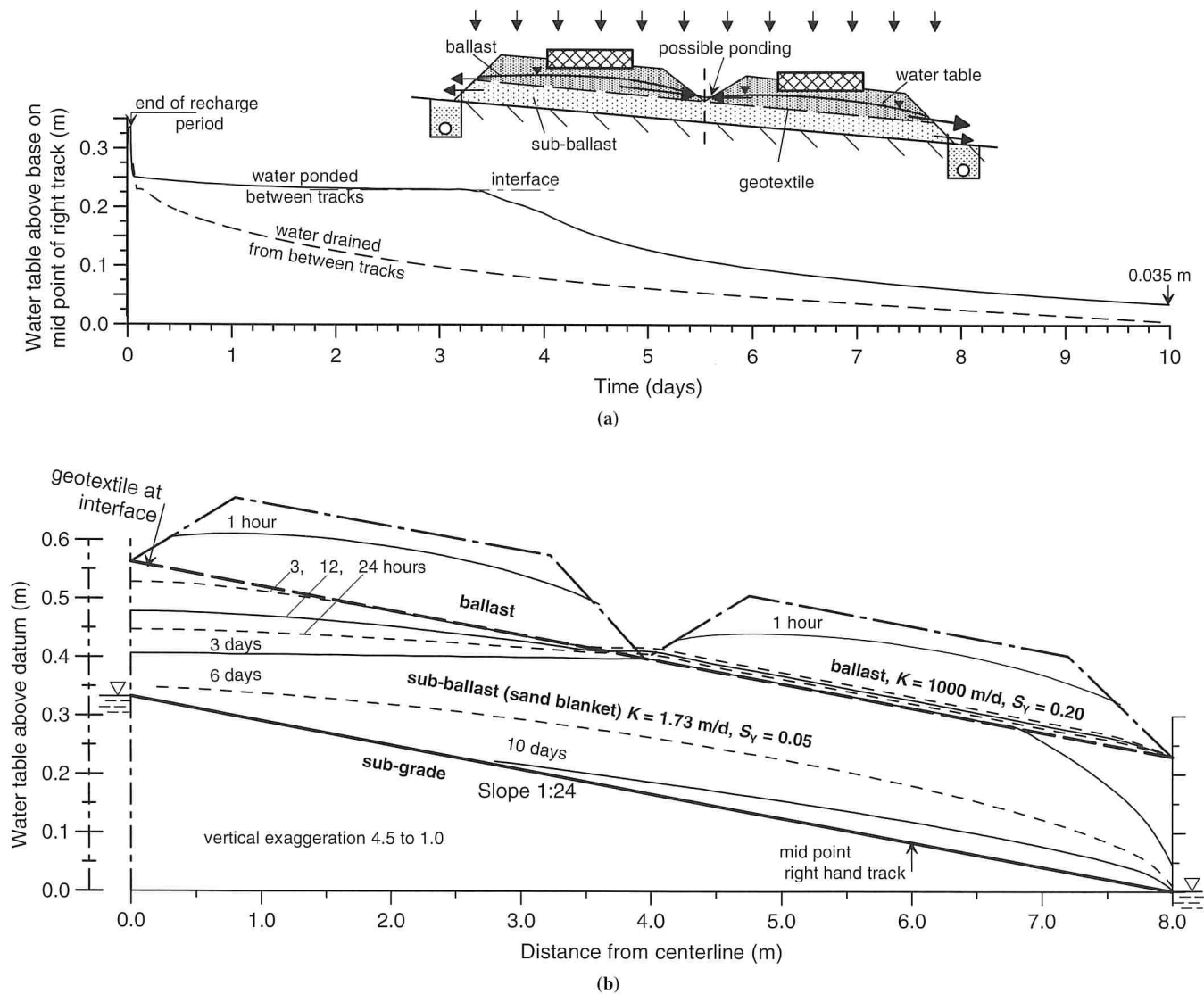


FIGURE 7 Representative problem for superelevated double track: (a) height of water table above subgrade at midpoint of lower track, solid line indicates water collects between tracks, broken line indicates water drains from between tracks; and (b) water table elevations on cross section when water collects between tracks (parameter values shown in Figure 9).

the permeability of the subballast is reduced, the time to reach the reference head of water above subgrade (1 cm) increases. In addition this result shows a trend similar to that for single track without any drainage geocomposite (see Figure 5).

The design curves shown in Figures 5 and 10 were developed by using a numerical model formulated with results of an experimental study undertaken by Heyns (4). Heyns's experiments also included drainage from alternative subballast with permeability of about 30% of the original material; the times for drainage were consistent with the design line of Figure 10.

### HIGH-PERMEABILITY GEOCOMPOSITES

For the third problem four possible arrangements of high-permeability geocomposites are examined (Figure 11). In the first arrangement there is no geocomposite, in the second the geocomposite is located between the subballast and the ballast, in the third the geocomposite is

placed in a sand blanket 3 cm above the subgrade, and in the fourth arrangement the permeable geocomposite lies on the subgrade. The permeability of the geocomposite is half the permeability of ballast (the actual value of the geocomposite permeability does not matter too much provided that it is within half an order of magnitude of the ballast permeability).

The location of a permeable geotextile between the subballast and ballast has a negligible effect on the reference time (4.3 days compared with 4.4 days). If the geocomposite is 3 cm above the subgrade, the time for the water table to fall within 1 cm at the midpoint beneath a track is 2 days, less than half the time for the first two arrangements. However, drainage from the subballast is far more rapid when the geotextile is immediately above the subgrade; the reference time is about 5% of that with no geocomposite.

A high-permeability geocomposite that may be used in such an application was described by Ghataora and Burrow (13). It comprises a needle-punched textile sandwiching a holed cusped layer. Thus, the composite exhibited high lateral and vertical permeability.

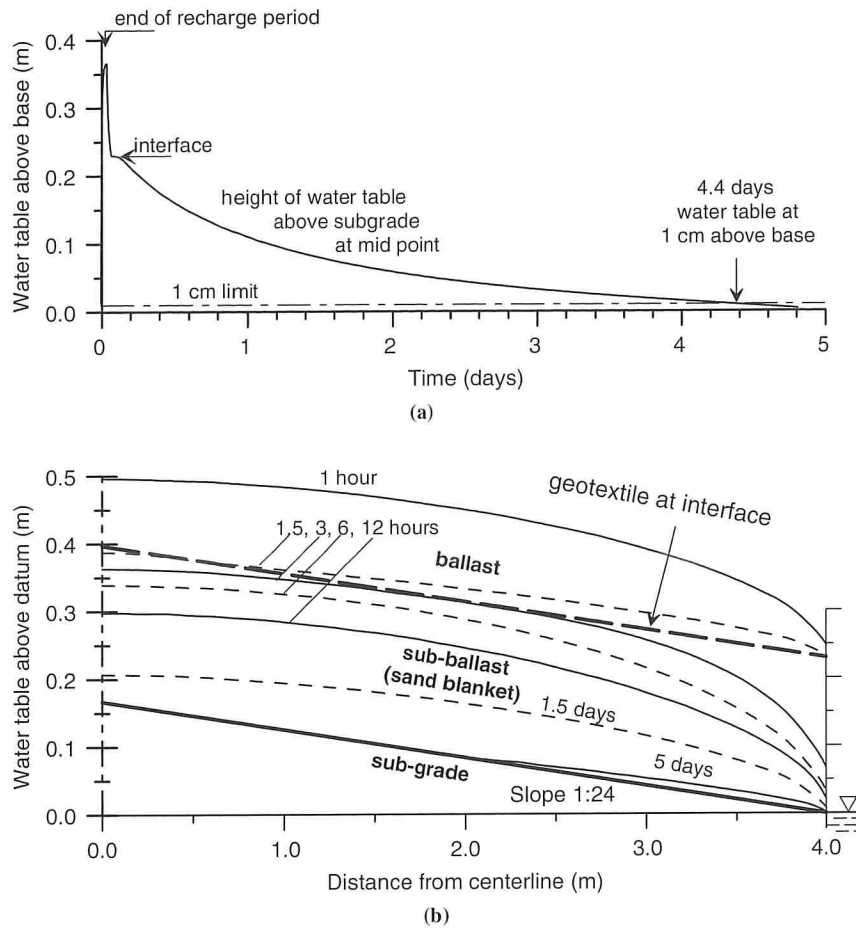


FIGURE 8 Representative problem for one side of symmetrical double track: (a) height of water table above subgrade at midpoint of one track and (b) water table elevations on cross section [parameter values: recharge = 3.05 m/day for 1 h; slope of subgrade 1:24; subballast 0.23 m thick, permeability 1.73 m/day ( $2.0 \times 10^{-5}$  m/s), specific yield = 0.05; ballast, permeability 1,000 m/day ( $1,157 \times 10^{-5}$  m/s), specific yield = 0.20].

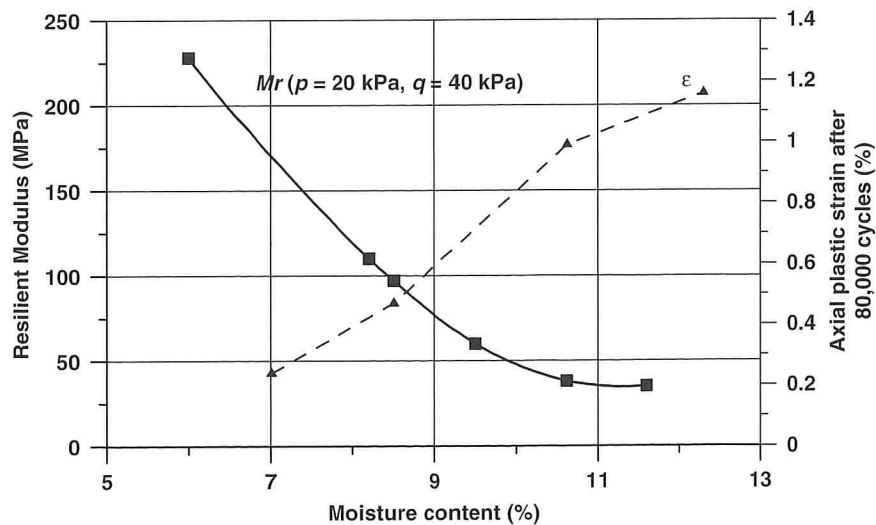


FIGURE 9 Effects of increase in moisture content on resilient modulus and plastic strain (11).

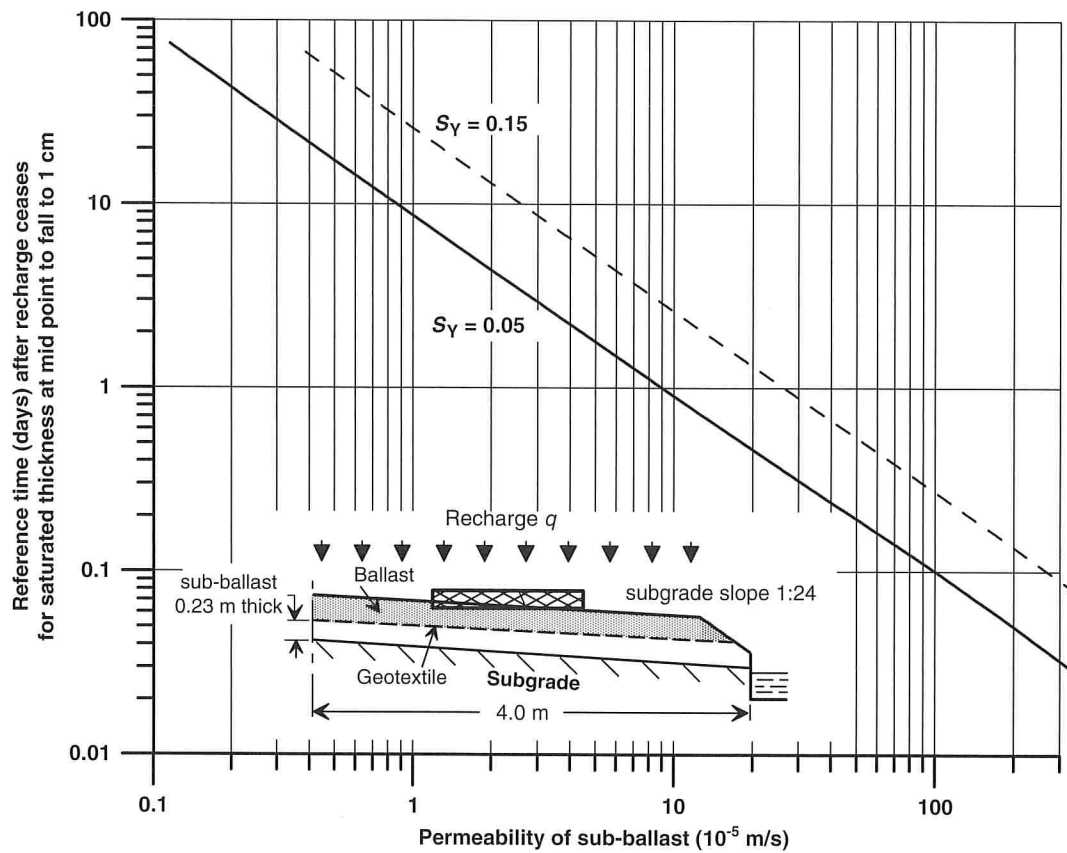


FIGURE 10 Design chart showing how permeability of subballast determines reference time for water table at midpoint of one track to fall to within 1 cm of subgrade with results plotted for specific yield of subballast of 0.05 and 0.15 [parameter values: recharge = 3.05 m/day for 1 h; slope of subgrade 1:24; subballast 0.23 m thick; ballast, permeability 1,000 m/day ( $1,157 \times 10^{-5}$  m/s), specific yield = 0.201.

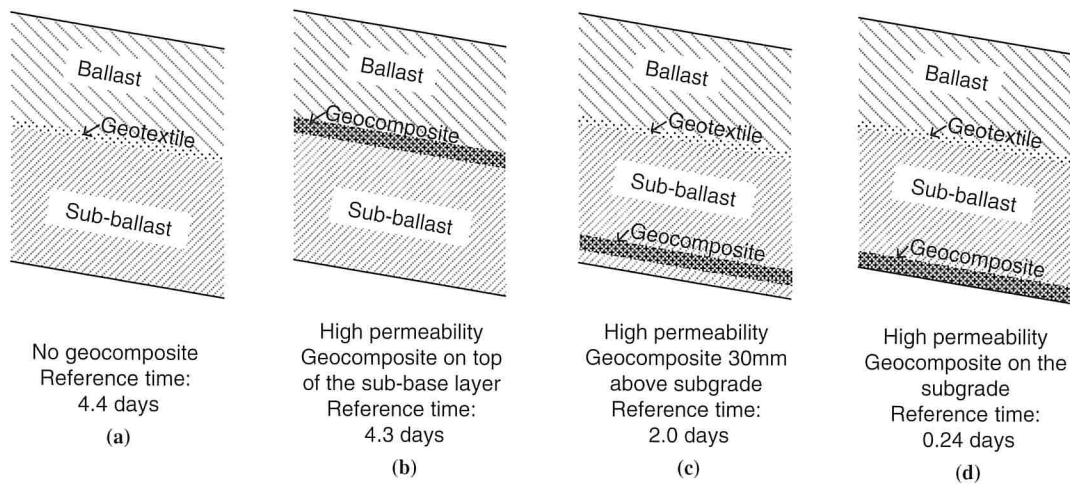


FIGURE 11 Comparison of reference times with respect to inclusion and position of permeable geocomposite (10 mm thick).

## CONCLUSIONS

This study is focused on the drainage of ballast and subballast for the case of dual railway lines for a typical scenario in the United States. The study's following findings are based on a numerical model that replicates Heyns' experiments (4):

- Water can take more than 4 days to drain from a single track. However, the inclusion of a high-permeability geocomposite near the base of the subballast can significantly improve the rate of drainage.
- The specific yield and permeability of the subballast affects the rate at which water drains out of the track.
- For the case of a dual superelevated track, if drainage is on one side only, then water may remain in the track for 10 days or more. Water is retained for a significantly longer period in the track nearest the drain. This suggests that the track near the drain, in dual railway track with a single side drain, is likely to show more deterioration as a result of prolonged water residence time.
- Dual-track drainage may be improved substantially by incorporating a high-permeability track composite near the base of the subballast.

## REFERENCES

1. Selig, E. T., and J. M. Waters. *Track Geotechnology and Substructure Management*. Thomas Telford, Ltd., London, United Kingdom, 1994.
2. Hyslip, J. P., and W. T. McCarthy. Substructure Investigation and Remediation for High Tonnage Freight Line. 2000. [www.arena.org/files/library/2000\\_Conference\\_Proceedings/00052.pdf](http://www.arena.org/files/library/2000_Conference_Proceedings/00052.pdf).
3. Ghataora, G. S., B. Burns, M. Burrow, and H. Evdorides. Development of an Index Test for Assessing Anti-Pumping Materials in Railway Track Foundations. *First International Conference on Railway Foundations*, University of Birmingham, Birmingham, United Kingdom, 2006, pp. 355–366.
4. Heyns, F. J. *Railway Track Drainage Design Techniques*. PhD dissertation. University of Massachusetts, Amherst, 2000.
5. Youngs, E. G., and K. R. Rushton. Dupuit–Forchheimer Analyses of Steady State Water Table Heights due to Accretion in Drained Lands Overlying Undulating Sloping Impermeable Beds. *Journal of Irrigation and Drainage Engineering*, Vol. 135, No. 4, 2009, pp. 467–473.
6. Youngs, E. G., and K. R. Rushton. Steady State Drainage of Two-Layered Soil Regions Overlying an Undulating Sloping Bed with Examples of the Drainage of Ballast Beneath Railway Tracks. *Journal of Hydrology*, Vol. 377, 2009, pp. 367–376.
7. Rushton, K. R., and G. Ghataora. Understanding and Modelling Drainage of Railway Ballast. *Proceedings of the Institution of Civil Engineers, Transport*, Vol. 162, 2009, pp. 227–236.
8. Rushton, K. R., and G. Ghataora. Design for Efficient Drainage of Railway Track Foundations. *Proceedings of the Institution of Civil Engineers, Transport*, 2012 (accepted for publication).
9. *AREMA Manual of Railway Engineering*. Chapter 1. American Railway Engineering and Maintenance-of-Way Association, Lanham, Md., 2007.
10. Five to 60 Minute Precipitation Frequency for Eastern and Central United States. Technical Memorandum NWS HYDRO-35. National Oceanic and Atmospheric Administration/National Weather Service, 1997.
11. Hornych, P., O. Hameury, and J.-L. Paute. Influence de l'eau sur le comportement mécanique des graves non traitées et sols supports de chaussées (in French). In *Symposium International AIPCR sur le Drainage des Chaussées*, Granada, Spain, 1998, pp. 249–257.
12. Hunt, G. A. *Optimisation of Track Formation Stiffness*. Report 1. British Rail Research Track Mechanics and Systems, London, 1993.
13. Ghataora, G. S., and M. P. Burrow. Composites at Ballast/Subgrade Interface in Railway Track Foundations. Presented at Railway Engineering–10th International Conference, London, June 24–25, 2009.

---

*The Railroad Track Structure System Design Committee peer-reviewed this paper.*

# Source of Ballast Fouling and Influence Considerations for Condition Assessment Criteria

Ted R. Sussmann, Mario Ruel, and Steven M. Chrismer

Railway ballast is a critical element in the railway track support structure. The ballast is often overlooked when inspection tools are developed for track. When ballast is not functioning correctly, the strength of the track structure may be inadequate and thus compromise track stability. Track stability-related failures vary from rapid deterioration with little warning to slow and progressive deterioration with often predictable required maintenance. Ballast-related deterioration is progressive and usually provides visual evidence to warn maintenance personnel of needed rehabilitation. However, the blocked drainage that develops with fouled ballast can result in a saturated roadbed that is not stable and could rapidly deteriorate to an unsafe condition with little warning. Although massive failures are rare, if a side hill fill or embankment deteriorates to the point of becoming susceptible to massive failure, then the challenge becomes evaluation. More detailed knowledge of the track support condition will be needed for a thorough evaluation than can be provided by current track inspections, except for costly detailed visual inspections. The current standard of practice for ballast inspection and maintenance can be improved to reduce the risk of sudden failure. Much of the required technology, knowledge, and resources is already available and being utilized under the current system. A more precise evaluation of ballast condition is essential to identify thresholds related to unsafe track support conditions and to support effective maintenance plans.

The main required ballast function pertaining to safety is that ballast must resist the forces applied to the track and retain the track in the required position (1, 2). Desirable ballast is angular stone material with a nominal size of approximately 0.75 to 2 in. and a uniform grain size distribution. Desirable ballast material is strong, hard, and durable crushed rock (aggregate) that does not readily deteriorate from applied loading, vibration, or environmental conditions or variations. This stone material will form a layer below and around the sides of the tie that provides direct vertical support and lateral stability. In this layer, the interparticle contacts between ballast particles will be points at which each angular particle interacts with an adjacent particle to transfer applied load. These small contact points deform

substantially under load to provide the high level of resilience common to clean ballast layers. The high loads involved and the small area of contact cause each ballast particle to deform substantially, potentially resulting in the fracture of particles or wear at particle contacts when the applied stress approaches or exceeds the strength.

Ballast resilience is obtained through the interparticle ballast contact points that provide for substantial vertical and lateral elastic deformation as long as the ballast stress remains tolerable and the ballast does not break down. The desired resilience in a ballast layer is elastic, meaning that the deformation is recoverable and the ballast will return to the position it occupied before loading. Deformation of fouled ballast is most often inelastic, meaning the deformation is often only partially recovered with significant plastic (permanent) deformation. In these locations the plastic ballast deformation results in track settlement and track geometry defects. According to Selig and Waters the ballast layer is generally the largest source of track settlement (2). Ballast-related track settlement can affect as little as one tie in the initial stages, and short wavelength track geometry parameters such as cross level, short wavelength profile, or short wavelength warp can be the initial indication of track support problems. As the track is repeatedly loaded, another inelastic deformation mechanism may occur in the ballast: the process of ballast compaction. Ballast compaction is sometimes referred to as “consolidation” in track stability-related publications and occurs where adjacent particles are forced ever more closely together under additional loading cycles from passing trains. Consolidation and compaction form a tight packing of ballast particles with increasing compressive contact force between particles to retain the particles in position as the compaction process continues with applied load cycles. The void space between the stones remains relatively large, which makes this layer very permeable, providing good drainage. The large interparticle compression force results in a highly stable ballast structure that retains track position well (2).

As the ballast progressively becomes fouled (contaminated with sand and silt-clay size particles), the void space between particles fills, gradually reducing the hydraulic conductivity and making the layer susceptible to rapid moisture-related deterioration. When the void space in the ballast layer becomes filled with fouling material, the necessary free movement of adjacent particles is limited and the compaction and consolidation process does not take place. As ballast fouls, the void space becomes partially occupied with fouling material that may temporarily increase the strength and stability of the layer. However, reduced drainage capacity and repeated loading of highly fouled ballast often cause the ballast particles to be forced apart by the fouling material either by mechanical movement of the particles under load in contact with fouling material or because the

---

T. R. Sussmann, U.S. Department of Transportation, Volpe Center, 55 Broadway, Kendall Square, Cambridge, MA 02142. M. Ruel, Canadian National Railway Triage Taschereau, Tour M, 1er Etage, 8050 Boulevard Cavendish, Montreal, Quebec H4T 1T1, Canada. S. M. Chrismer, Amtrak, 30th Street Station, Philadelphia, PA 19104. Corresponding author: T. R. Sussmann, ted.sussmann@dot.gov.

*Transportation Research Record: Journal of the Transportation Research Board*, No. 2289, Transportation Research Board of the National Academies, Washington, D.C., 2012, pp. 87–94.  
DOI: 10.3141/2289-12



fouling material reduces the friction at interparticle contact points, which releases some of the interparticle compressive force, thus reducing the stability of the ballast and the entire track structure. Fouling material is often most noticeable as muddy slurry. The fouling material and slurry most often consist mainly of ballast breakdown. However, either subgrade migration into the ballast layer or ballast into the subgrade may result in the presence of subgrade soil in the fouling material. Subgrade may also appear as a result of attrition in which the hard ballast layer is placed directly on a soft rock subgrade (e.g., clay shale) and the applied stress at the ballast subgrade interface results in wear of the subgrade surface, the result of which can be pumped into the ballast when it is saturated. Other foreign materials such as blown sand and coal dust can also be found in fouling material. Fouled ballast and the typical muddy slurry are difficult to inspect with common methods. The methods along with the reasons for the inspection difficulties follow:

1. Visual inspection because of opaqueness,
2. Electromagnetic methods because of typically high conductivity, and
3. Mechanical means of load deflection testing because of variability of test results.

Although the inspection of ballast is difficult, the role in track performance is critical and techniques are advancing for systematic evaluation of ballast condition. This paper describes the processes involved in ballast performance and failure and identifies the critical stages and methods for assessment.

## BALLAST COMPACTION AND CONSOLIDATION

The ballast compaction and consolidation process is the densification process occurring after new ballast is placed or existing ballast is tamped. The loose ballast that results from ballast placement or tamping is often compacted under traffic or with a dynamic track stabilizer. Initial densification of loose ballast requires either a track stabilizer or a period of slow traffic to consolidate the ballast and mitigate the risk of buckling while the track lateral resistance is at its lowest point (3). In track lateral stability analysis, the change in lateral strength of ballast over time and with traffic during ballast compaction has been called consolidation to refer to the increase in overall track strength with time. However, this is technically a compaction process of the ballast. The density increase results in high interparticle contact force that holds the ballast in position with increased lateral track strength common to consolidated track.

The ballast compaction process starts with the loose ballast shown in Figure 1a. The loose ballast contacts adjacent particles at discrete locations, but over a small area such that in any cross-sectional slice, the contact point may not be seen, as in this figure. As either the dynamic track stabilizer or passing traffic compacts the ballast, the particles are forced into a tighter packing, which increases the contact area between particles (Figure 1b). Further traffic loading will further force particles into a tighter packing (Figure 1c), where Particle A interlocks with Particles C and D because Particle B forces Particle A between those particles. The asperities on Particles A and C interlock, which results in a strong response to load and often which results it difficult to remove this particle from the ballast matrix.

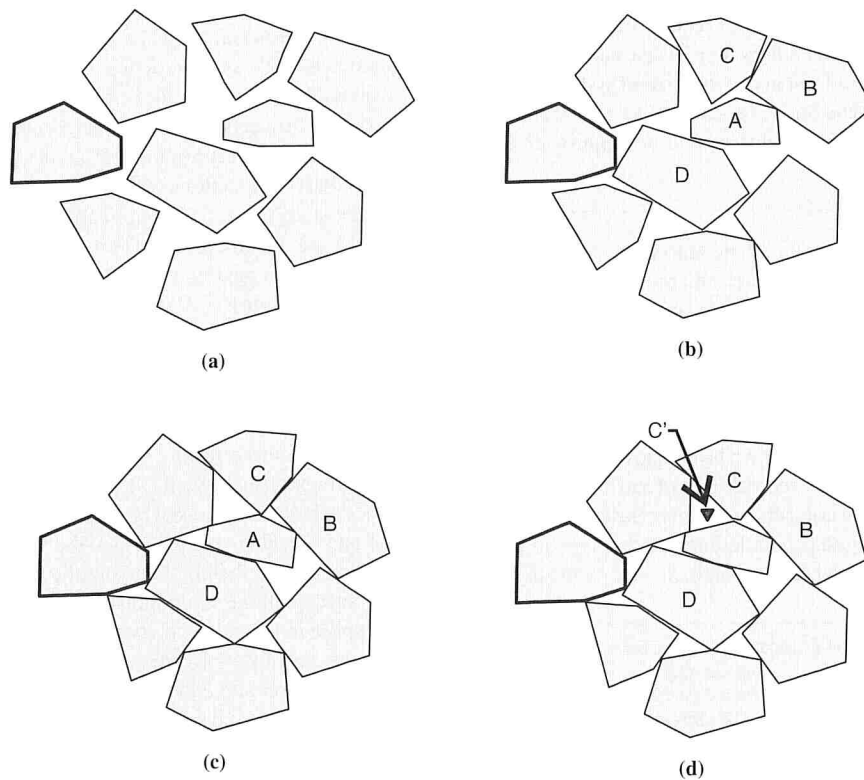


FIGURE 1 Stages of ballast compaction and breakdown: (a) new ballast placement, (b) initial ballast compaction, (c) well-compacted ballast typical of consolidated track, and (d) ballast breakdown.

On further loading, the ballast could continue to compact depending on the strength of the ballast and the magnitude of the applied load. If the ballast is adequately strong to support the applied loads, then the ballast compaction process proceeds with some ballast abrasion and breakage to develop the tight ballast matrix common to strong, consolidated track. If the applied loads exceed the ballast particle strength, the ballast will break down. In Figure 1*d*, the asperity of Particle C breaks off and forms Particle C' as a result of the applied load that exceeds the particle strength. As the particle contact forces approach the ballast particle strength, further increased loading likely requires either ballast particle movement to shed the applied load on Particle C to neighboring particles or particle breakage.

In regard to ballast contact area, the particle-to-particle contact area for the ballast material is a useful point of reference to consider variations in ballast structural stability. As illustrated in Figure 2, the contact area varies through the same four stages shown in Figure 1. The ballast-to-ballast particle contact area varies from when the ballast is freshly placed and loose (Figure 1*a*) to increased particle contact as the ballast rearranges and reorients into a more stable structure on initial compaction (Figure 1*b*). The consolidated ballast is the desired condition in which ballast can remain in a stable orientation for an extended period of time and traffic (Figure 1*c*). Further loading can deteriorate the ballast to the point at which the particles degrade and the voids fill. As the void space fills, the ballast structural matrix that has carried the load and distributed the load to the lower layers begins to shed load to the weaker material in the void space as the ballast particles are forced apart by the fouling material and the ballast structural performance degrades. Maintenance intervention will not substantially change this figure because ballast tamping will result in completely loosened ballast that will need to recompact and consolidate (Figure 1*d*) if a stable, consolidated ballast layer is to develop.

The ballast consolidation process can proceed from loose ballast to dense with continued compaction under applied load to develop a very strong and stable foundation for the track. The continued loading of the stable, compacted ballast may result in very little additional ballast particle movement after a stable ballast structure has developed. Once the ballast has densified, any applied load that exceeds the historic range of loadings will cause the ballast rearrangement process to restart. In cases in which interparticle contact

forces exceed particle strength, ballast will break and abrade to support and distribute the applied loads.

Water that is trapped in the track by the fouling material will be subject to the high stress of passing traffic and result in the pumping of ballast-fouling material to the track surface. The transfer of transient traffic loading-related stress to the trapped water is common and results in ballast mud spraying up under passing traffic and the formation of mud boils where the fouling material exits the track structure under pressure. The mud from the ballast breakdown, water, and any other ballast-fouling material will result in less resistance to ballast rearrangement by lubricating the ballast-to-ballast particle contacts, often resulting in track settlement and geometry deviations.

## BALLAST-FOULING MATERIALS

Ballast-fouling material is most often the result of ballast breakdown under load. The fouled ballast problem areas that are most often investigated are muddy fouled ballast zones that contain gravel from ballast breakdown combined with fine-grained clay and silt particles that result from ballast deterioration or that may fall or blow onto the track. Silt and clay can also pump into the ballast from lower layers of old fouled ballast or, in a relatively small number of cases, from the subgrade where subballast or another layer does not exist to act as a filter to separate the subgrade from the ballast. However, the plastic nature of common clay subgrades causes the material to tend to stick together, making it less common that the particles will separate and pump up into the ballast except when saturated. Most often the ballast has been pushed down into the subgrade in instances in which subgrade is present in track with ballast. The behavior of fouled ballast is often dominated by even a small percentage of silt and clay that can create a fouling material that is plastic with a lubricating quality that will limit particle interlocking. Once the ballast is adequately fouled so that the ballast particles can no longer interact without mobilizing the weak structural response of the fouling material, the ballast function is compromised. At this point, the ability to retain track geometry under even a few load cycles will be impossible (2).

Ballast breakdown results in smaller particles of ballast and, when not contaminated with other detritus, this fouling material sometimes has the consistency of gravel that might increase the strength of the ballast section. If it occurs, the increased ballast layer strength might be the result of strong ballast breakdown particles in the ballast voids that might restrain ballast particle movement, thus reducing track resilience but potentially increasing strength and structural stability. However, there is a fine line between strengthening the ballast with breakdown material and inhibiting drainage. Data on locations where ballast stability is enhanced are lacking because investigations of well-performing track sections are often not completed. However, many examples exist to demonstrate the wide variety of problems related to ballast fouling. In general, small changes in ballast void opening size can significantly affect drainage characteristics and would more likely limit ballast life-cycle performance.

Coal fouling is much like fine-grained soils (silt and clay), except that the detrimental characteristics of high affinity for water, low strength, low stiffness, and high plasticity can be more pronounced in some types of coal. Assuming that the source of coal in the track structure is from debris that has blown or fallen onto the track, train traffic will vibrate the coal in and between particles. Unlike many other ballast-fouling processes, the contamination of ballast with coal

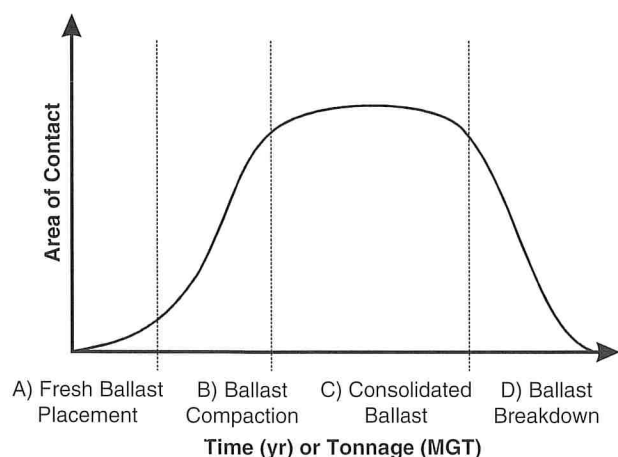


FIGURE 2 Ballast-to-ballast particle contact area change over time or tonnage.

can occur before the ballast begins to break down. The result is the unusual condition in which the fouling material can be nearly 100% coal. Working from the top down, the coal fouling material first affects the strength in the upper ballast where much of the resistance to applied loading is developed. Critically, it is this zone that provides the most crucial elements of stability: resisting both lateral and vertical movement. Weakness in the upper ballast can compromise lateral resistance and result in increased probability of track buckling resulting from thermal or other rail compressive forces or track shift under repeated loads. As it works down into the ballast section under train loading, the coal will reduce vertical load support.

Ballast-fouling material is commonly measured, reported, and compared by weight in cases in which experience has been used to establish thresholds. In the case of coal or any other relatively low specific gravity material, the same weight of material will occupy a larger percentage of the void space in ballast compared with traditional fouling materials. Considering the common specific gravities for coal (0.8–1.3), granite (2.65–2.75), and clay (2.3–2.6), the same weight of coal will generally occupy about two times the volume of either clay or granite as a result of the lower specific gravity. This large volume means that half the weight of coal dust will be required to occupy the same void space and disrupt ballast performance (such as drainage) to the same extent as common fouling material. Furthermore, the available experience-driven thresholds of acceptable amounts of ballast fouling are all based on weight and will not be valid for fouling materials with a lower specific gravity.

Selig and Waters used the fouling index (FI) to account for the rapid track deterioration common with ballast fouling from finer material such as silt and clay that characteristically passes the No. 200 sieve (2). Because of the influence of clay and silt in reducing ballast performance, Selig and Waters proposed the  $FI = \text{percent by weight passing the No. 4 sieve} + \text{percent by weight passing the No. 200 sieve}$  (2). (The numbers 4 and 200 refer to holes per inch of screen that make up the sieve.) This double-counts the fine material passing the No. 200 sieve (since the fine silt and clay pass both the No. 4 and the No. 200) because of its large influence on performance deterioration.

The FI will not provide consistent results for coal dust because of the density difference of the coal fouling material, which will result in an artificially low FI. A factor on the order of two, but dependent on the specific gravity of the main fouling materials, would be needed to adjust the FI to account accurately for density differences. This adjustment would not include any additional detriment to the ballast structure resulting from the particular detrimental qualities of the coal dust that might inhibit performance.

Several common gradations of fouled ballast are presented in Figure 3. Clean ballast is the gradation curve to the bottom left, with increasing FI up and to the right. The change in gradation curve is significant in that a tail of the curve develops where smaller particles accumulate in the ballast by wear and contamination. Selig and Waters also qualify the ranges of FI according to Table 1 (2).

## DRAINAGE

One of the main performance limitations for fouled ballast is the requirement that fouled ballast must freely drain, which is impossible when the voids become full (2). On the basis of lab experiments to determine the critical rainfall rate at which the track will saturate as the precipitation flows from the track centerline to the ditch, Selig and Waters developed the data presented in Figure 4 (2). These data show the influence of ballast fouling on the reduced drainage capac-

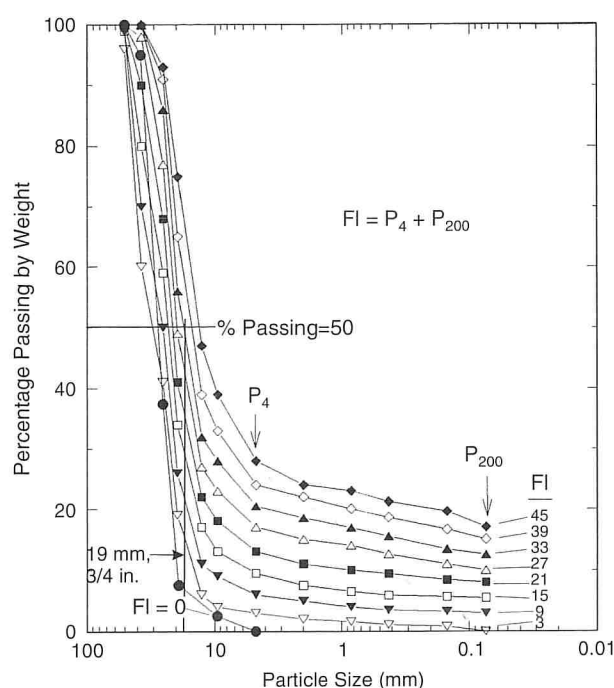


FIGURE 3 Influence of fouling on ballast gradation curve [modified according to figures in Selig and Waters (2)].

ity of ballast. As the ballast voids become full, the amount of flow between particles and in the void space is reduced, thus trapping water in the track structure and potentially saturating the track. On the basis of these data, an FI of 30 is the breakpoint at which track drainage becomes substantially more impeded by fouling.

## MECHANICS OF FOULED BALLAST

As ballast-fouling content increases, a point can be reached at which the void space is full and any additional increase in fouling material necessitates that the ballast particles move apart to accommodate the fouling material. In these cases, the ballast-fouling material that fills ballast voids results in ballast strength and stiffness that will more resemble the behavior of the relatively weak fouling material with a weak strength and stiffness response to applied loading, at least at small deformations. At small deformations the resistance will be dominated by the soft response of whatever fouling materials

TABLE 1 Assessment of Ballast Fouling and Fouling Index

Category	Fouling Index (2)	Percent Passing 3/4-in. Sieve (CN)
Clean	<1	—
Moderately clean	1 to <10	—
Moderately fouled	10 to <20	25–35
Fouled	20 to <40	40–50
Highly fouled	>40	>50

NOTE: CN = Canada National Railway; — = not applicable.

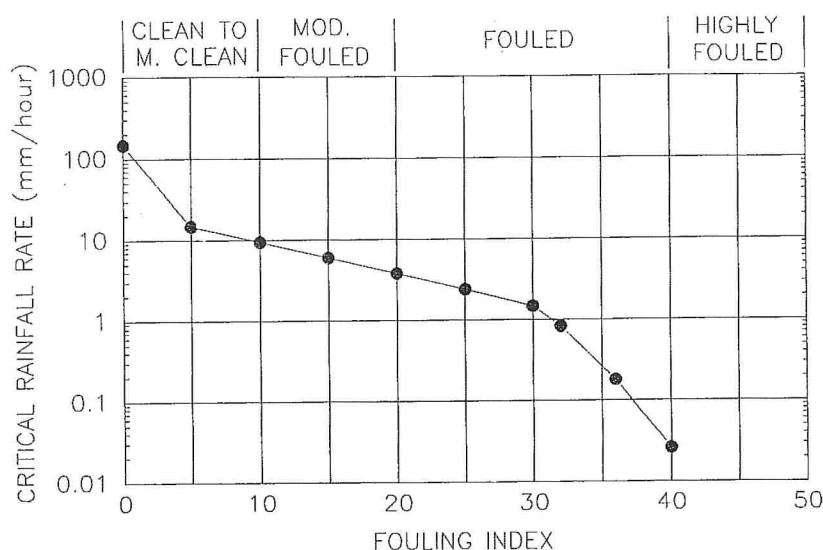


FIGURE 4 Critical rainfall rate and fouling index variation (m. = moderately; mod. = moderately) (2).

exist between particles. At larger deformation, the ballast particles will again be forced into contact, to resist the applied loads. However, the resistance offered will be much lower than anticipated because the ballast will not be consolidated—like recently tamped track. However, unlike with recently tamped track, the ballast particles will be surrounded by fouling material, which will provide an even weaker response than does loose, clean ballast.

For newly constructed track, the criterion for proper placement of ballast has been assumed to be the ballast density at which a minimum strength could be determined to correspond to a minimum strength of the ballast to resist applied loads. However, the challenge has been to develop a ballast density measurement that can be effectively applied. The size of the ballast particles requires specialized equipment that is cost prohibitive. However, new concepts in the measurement of the stiffness of placed materials and track deflection testing provide the opportunity to measure track response to load in a cost-effective manner in which a more direct measure of the ballast stability to support applied traffic loading may be obtained (4). In this manner, the construction specifications would control track deflection and could be verified by a deflection test before track acceptance in which loose ballast would be identified by zones of increased deflection.

During construction such a specification should be supported by new intelligent compaction technology in which the response of the compactor is used to determine whether the material stiffness has increased adequately to meet density specifications. However, density specifications developed because direct measurement of the strength and stiffness required of soil has not been practical. Intelligent compaction solves this problem by providing a direct measurement of the stiffness of the soil based on the response of the vibratory compactor. Chang et al. describe the current state of the technology for aggregate compaction, which could be extended to include ballast aggregate (5). This technology could be implemented as described in Chang by using traditional compaction equipment; however, deployment in a more integrated manner as part of a track laying machine or other on-track equipment would improve productivity (5).

Coal dust fouling of ballast provides a unique problem because coal dust contaminates ballast from the top down, where the applied loads are the highest and ballast stability is critical. Anecdotal reports

suggest that coal can accumulate in ballast in two ways: coal will either coat and surround the ballast particles or accumulate into a fairly sizable mass. The colloidal behavior of fine coal dust could lead to attraction of coal to ballast and to coal dust coating the ballast particles. In addition, masses of coal dust have been reportedly located in track. In locations with severe coal dust fouling, zones where ballast voids are completely filled with ballast may appear to be a solid mass of coal that may have fallen from a passing car. Regardless of whether coal coats ballast or accumulates as a mass to fill the void space, the reduction in ballast strength could present additional maintenance challenges and safety inspection requirements.

The common range for ballast-fouling material is based on the concept of available void storage volume. Open-graded, large angular crushed rock provides a relatively large void space. On the basis of typical ballast properties of unit weight of 100 lb/ft<sup>3</sup> and specific gravity of 2.7, the void space is roughly 41% of the volume of the ballast (Table 2), although these estimates are heavily dependent on particle shape and gradation (7). As voids become progressively filled with fouling material, the mobility of adjacent ballast particles under applied load will be limited, the ballast particles will interact with the fouling material and be forced apart, or the fouling material may lubricate ballast contact points. In either case the ballast will lose the interparticle contact force common to consolidated ballast.

The Canadian National Railway (CN) has followed ballast performance for several years and tracked the change in gradation

TABLE 2 Unit Weight–Volume Relationships of Ballast at Several Stages (6).

Ballast Condition	FI	Unit Weight (lb/ft <sup>3</sup> )	Void Ratio	Void Volume (%)
Loose	0	100	0.68	41
Consolidated	0	110	0.53	35
Fouled	20	125	0.35	26
Failing	40	135	0.25	20

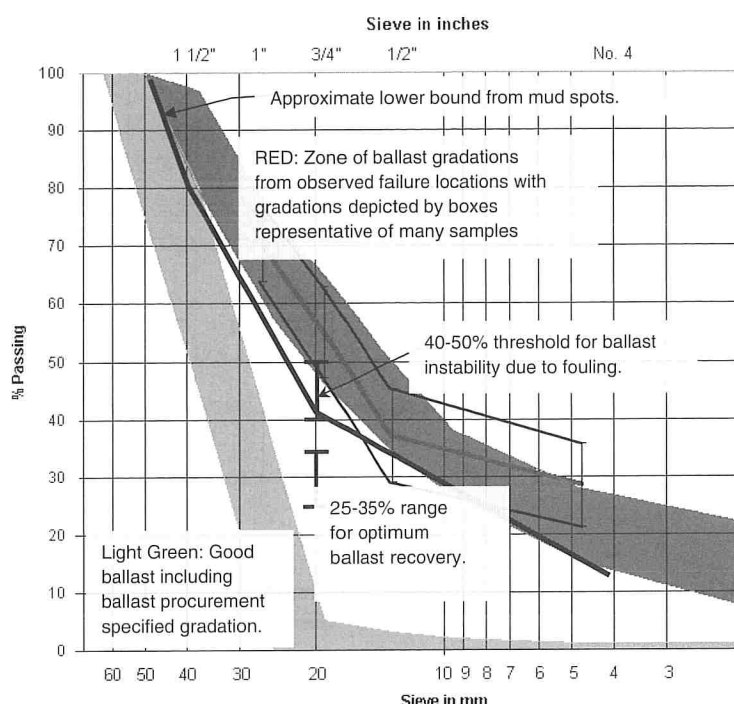


FIGURE 5 Gradation curve zones of observed adequate and inadequate ballast performance.

curve over time as ballast breaks down. Figure 5 shows the light green zone of acceptable ballast performance from new ballast specification to initial wear of ballast through the consolidation process and then for approximately 500 million gross tons of traffic where the red zone depicts the range of gradations for fouled ballast. The wear of the ballast progressed from the point that nearly all the ballast was larger than the  $\frac{3}{4}$ -in. (19-mm) sieve to locations of observed ballast instability where the ballast had worn to the point that approximately half the material was  $\frac{3}{4}$  in. (19 mm) or smaller. From this study, the specification of undercutting is targeted to 25% to 35% passing the  $\frac{3}{4}$ -in. (19-mm) sieve because the progression of ballast breakdown can be rapid as the voids become full and drainage is impeded. These limits are compared with FI thresholds from Selig and Waters in Table 1 to relate the two criteria (2).

The zone of gradations from ballast failure locations in Figure 5 indicates that 50% or more passing the  $\frac{3}{4}$ -in. (19-mm) sieve falls in an FI range of 33 to 39, according to the gradations in Figure 3. This result corresponds well with the limits set by Selig and Waters noted in Table 1, in which this percentage falls near the high end of fouled, approaching highly fouled (2). In fact, considering that CN did not measure the percent passing the No. 200 sieve for these tests, it is likely that the FI would have exceeded 40 and corresponded to highly fouled ballast if the percent passing the No. 200 sieve had been measured to compute the FI fully.

Further review of the data from CN indicates that about 30% passing the  $\frac{3}{4}$ -in. (19-mm) sieve is the threshold to undercut ballast (actually 25% to 35% in Figure 5). This figure corresponds to an FI of approximately 21 to 27 for the ballast gradations presented in Figure 3, which is close to the FI = 30 threshold noted for drainage-related considerations in Figure 4. For this gradation, the ballast unit weight was approximately 135 lb/ft<sup>3</sup> with an associated void volume of 20% (Table 2) according to Selig and Waters and Chang

et al. (2, 5). In this case the void volume has been reduced from 41% for the clean ballast to 20% of the entire ballast volume (Table 2). According to the data in Figure 4, the ballast unit weight was approximately 125 lb/ft<sup>3</sup> at FI = 30, at which the void space is approximately 35%. This analysis gives a range of void volume and expected behavior in Table 2 in which clean, loose ballast has a void volume of 41%, working ballast has a void volume range of 26% to 35%, and failure could be expected at void volumes on the order of 20%, at which about half of the initial void volume has been filled with fouling material.

## FOULED BALLAST MEASUREMENT

The fouling of ballast is progressive and can be influenced by many factors. In cases in which no environmental or operational detritus contaminates the ballast, ballast breakdown can be expected and may initially enhance stability by retaining ballast particles in place. However, this is a temporary condition and either damage caused by partially blocked drainage or continued wear of the ballast will occur to destabilize the ballast. In cases in which blown, dropped, pumped from below, or mixed material from other layers contaminates the ballast, it is important to note the influence and develop parameters to quantify the influence such as the FI or the CN  $\frac{3}{4}$ -in. sieve-based threshold.

"Ballast breakage" is a term that generally refers to any damage from the initial desired gradation of ballast. This damage can be caused by placement, transport, or maintenance. One key point is that the smaller particles that develop due to ballast breakage are still relatively strong ballast particles that may be able to distribute load reasonably well. For better tracking of ballast breakdown, the ballast breakage index suggested by Indraratna and Salim appears to



be useful (7). The ballast breakage index is the sum of the additional weight retained on each sieve as the ballast deteriorates and breaks down. This index is useful because the larger particles will wear and break, resulting in smaller particles, and the additional weight provided by these broken particles on the smaller sieves provides a direct indication of ballast performance. Equivalently, the breakage index could be the summation of all weight lost on the larger sieves because the weight lost due to particle breakage must equal the weight gain on smaller sieves. The advantage of this change is that used ballast is difficult to clean and sieve and the larger particles occupying the larger sieves present less of a challenge to the cleaning and sieving process.

In practice the initial gradation may not be available and local variations may be substantial, which makes the comparison from the initial state impossible. Even so, the general concept appears useful if a standard set of sieves were used for acceptance verification and field evaluation through the ballast life. By tracking the weight change for specific sieve sizes, particular ballast particle failures can be identified and characterized. When particles actually crush or split, the sieves that would gain the broken particles are different from those if the ballast were wearing due to abrasion. This concept may be critical in the evaluation of ballast stability. For example, Indraratna and Salim compared new and recycled ballast in laboratory testing to evaluate changes in stability (7). The new ballast was a standard gradation, whereas the recycled ballast had been worn in track. The data showed that clean ballast developed a high degree of particle interlocking at the low confining stress common in track, resulting in a high strength at failure. The recycled ballast did not interlock as well, presumably because the asperities required to develop a high degree of interparticle stability had been worn away. Ballast that crushed under load might continue to have the needed asperities for particle interlocking, even though the particles were smaller, whereas abraded ballast would not. A ballast breakage measurement could track those performance variations. However, sample variability and measurement challenges will likely limit the practical application of this measurement concept.

## BALLAST CONDITION ASSESSMENT

Two methods currently exist that could provide information on ballast performance: vertical track deflection and ground penetrating radar (GPR). Track deflection measurement has been developed by several groups including the Transportation Technology Center, Inc., and the University of Nebraska–Lincoln in the United States (8–10). The measurement of vertical deflection provides a method to identify risks associated with excessive deflection of the track and poor track support (9). A large track deflection along with variations in track deflection along the length of the track can be expected as a result of fouled ballast (8). However, the distinct performance differences of dry and wet fouled ballast present a challenge because dry fouled ballast may not deform any more than track with clean ballast whereas wet fouled ballast may deform substantially. Because ballast inspection cannot be effective if it is dependent on weather and the moisture condition of the track, other techniques can be used to augment the track deflection measurement and provide a more direct measure of ballast condition. However, it must be acknowledged that the vertical deflection of the track is the single best indicator of track life and should be applied in conjunction with a specific ballast evaluation measurement (11–13).

The leading technology for ballast condition assessment is GPR, which provides a complementary data set to track deflection that

will support a more complete evaluation of track condition. GPR is an electromagnetic inspection technique in which a pulse of electromagnetic energy is transmitted into the track structure and the resulting reflections are recorded and interpreted in relation to track structure-supporting materials and layers (14, 15). The lack of automated data interpretation is a hindrance to more widespread use of this technique, but for relatively shallow layers such as the ballast it is possible to develop automated inspection algorithms that simplify interpretation and make the technology economically feasible (16–18). In these interpretation schemes, the GPR signal return from the ballast is tracked and the interface between clean and fouled ballast turns out to be a key measurement indicative of track performance. As the fouled ballast depth affects tie support, track stability and performance deteriorate (16, 17). Typically clean ballast depths of at least an inch or so deeper than the bottom of the tie are needed to adequately support the track.

## SUMMARY AND CONCLUSIONS

The stability of railway ballast is critical to the efficient performance of railway track. Ballast breakdown is inevitable under heavy load and heavy tonnage. However, when ballast is kept clean from contamination and drains freely, the ballast breakage that develops might temporarily enhance stability by contributing to the interlocking of adjacent particles. However, the breakdown will occupy void space and potentially inhibit drainage and reduce ballast life. When ballast breakdown becomes contaminated with other ballast-fouling material, the common muddy fouled ballast problem develops as a result of the presence of silt and clay that can blow or wash into the track structure or be pumped or mixed from other track structural layers. Although typically only representing a small percentage of fouling material compared with ballast breakage, the presence of silt and clay fouling materials dominates the performance by lubricating the particles and reducing particle interlocking. Coal dust can further exacerbate these problems as a result of even weaker structural response, and the low specific gravity could lead to misleading estimates of the amount of fouling in the ballast. By tracking both ballast breakage and contamination, ballast performance may be better characterized. In practice, quick field tests such as the 25% to 35% retained on the  $\frac{3}{4}$ -in. sieve undercutting threshold developed by CN provide a quantitative means to evaluate distinct sites and gauge expected performance on the basis of previous experience. Application of track deflection or ground penetrating radar technology appears to be the most direct and available means to assess ballast condition and provide a continuous evaluation of thresholds. As these technologies mature, additional terminology will be required to identify track performance problems. At this stage, it may be possible to track the depth of clean ballast in track. As a start, identification of locations where the clean ballast extends below the bottom of the tie appears to be a useful tool to discriminate sites that may be expected to perform well from those sites that may not be stable and may require maintenance.

## ACKNOWLEDGMENTS

Specific thoughts on ballast performance, evaluation, and quantification have developed during many discussions of the Substructure Technical Advisory Group for Transportation Technology Center, Inc. (TTCI). The input of the entire committee has been critical,



especially from Dingqing Li and Dave Read of TTCI, who have provided leadership and motivation for the group.

## REFERENCES

- Hay, W. W. *Railroad Engineering*, 2nd ed. John Wiley and Sons, New York, 1982.
- Selig, E. T., and J. M. Waters. *Track Geotechnology and Substructure Management*. Thomas Telford, London, 1994.
- Sussmann, T., A. Kish, and M. Trosino. Influence of Track Maintenance on Lateral Resistance of Concrete-Tie Track. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 1825, Transportation Research Board of the National Academies, Washington, D.C., 2003, pp. 56–63.
- Farritor, S. *Real Time Measurement of Track Modulus from a Moving Car*. Report Number FRA/ORD-05/05. FRA, U.S. Department of Transportation, Dec. 2005.
- Chang, G., Q. Xu, J. Rutledge, B. Horan, L. Michael, D. White, and P. Vennapusa. *Accelerated Implementation of Intelligent Compaction Technology for Embankment Subgrade Soils, Aggregate Base, and Asphalt Pavement Materials*. Report Number FHWA-IF-12-002. FHWA, U.S. Department of Transportation, 2011.
- Yoo, T. S., H. M. Chen, and E. T. Selig. Railroad Ballast Density Measurements. *Geotechnical Testing Journal*, Vol. 1, No. 1, March 1978, pp. 41–54.
- Indraratna, B., and W. Salim. *Mechanics of Ballasted Rail Tracks: A Geotechnical Perspective*. Taylor and Francis, London, 2005.
- Li, D., R. Thompson, P. Marquez, and S. Kalay. Development and Implementation of a Continuous Vertical Track-Support Testing Technique. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 1863, Transportation Research Board of the National Academies, Washington, D.C., 2004, pp. 68–73.
- Carr, G. *Dynamic Response of Railroad Track Induced by High Speed Trains and Vertical Stiffness Transitions with Proposed Method of Measurement*. MS thesis. Tufts University, Medford, Mass., 1999.
- Sussmann, T. R., and E. T. Selig. Characterization of Track Substructure Performance. *Recent Advances in the Characterization of Pavement Geomaterials*. Geotechnical Special Publication, ASCE, 1999, pp. 37–48.
- Selig, E. T., and D. Li. Track Modulus: Its Meaning and Factors Influencing It. In *Transportation Research Record 1470*, TRB, National Research Council, Washington, D.C., 1994, pp. 47–54.
- Hunt, G. A. Track Damage Models as Tools for Track Design Optimization. *Proc., Conference on Innovations in the Design and Assessment of Railway Track*, Technical University of Delft, Netherlands, Dec. 1999.
- Ebersöhn, W. E. *Substructure Influence on Track Maintenance Requirements*. PhD dissertation. Geotechnical Report AAR95-423D. University of Massachusetts, Amherst, 1995.
- Sussmann, T. R. *Application of Ground Penetrating Radar to Railway Track Substructure Maintenance Management*. PhD dissertation. University of Massachusetts, Amherst, 1999.
- Sussmann, T. R., K. O'Hara, and E. T. Selig. Development of Material Properties for Railway Application of Ground Penetrating Radar. *Proc., SPIE Ninth International Conference on Ground Penetrating Radar* (S. K. Koppenjan and H. Lees, eds.), Vol. 4758, pp. 42–47.
- Roberts, R., I. Al-Qadi, E. Tutumluer, J. Boyle, and T. Sussmann. Advances in Railroad Ballast Evaluation Using 2 GHz Horn Antenna. Presented at International Conference on Ground Penetrating Radar, Columbus, Ohio, June 19–22, 2006.
- Roberts, R., I. Al-Qadi, E. Tutumluer, and J. Boyle. *Subsurface Evaluation of Railway Track Using Ground Penetrating Radar*. Report Number FRA/ORD-09/08. FRA, U.S. Department of Transportation, 2008.
- Silvast, M., M. Levomaki, A. Nurmikolu, and J. Noukka. NDT Techniques in Railway Structure Analysis. Presented at 7th World Congress on Railway Research, Montreal, Quebec, Canada, June 2006.

---

*The Railway Maintenance Committee peer-reviewed this paper.*

# Ground-Penetrating Radar Data to Develop Wavelet Technique for Quantifying Railroad Ballast-Fouling Conditions

Pengcheng Shangguan, Imad L. Al-Qadi, and Zhen Leng

Ballast fouling is detrimental to railroad track functions. Ground-penetrating radar (GPR), a nondestructive testing tool, has been used to assess ballast-fouling conditions. However, processing the extensive amount of data to quantify ballast conditions is challenging. Although several approaches have been developed, each of them has certain disadvantages, such as being unable to detect fouling without a clear interface between clean and fouled ballast and being user dependent. To overcome these disadvantages, a new approach based on wavelet transform was investigated. Laboratory tests were conducted to collect GPR data on ballast with controlled fouling levels. The data were processed by choosing the proper mother wavelet, selecting appropriate wavelet decomposition coefficients, and de-noising the signal. Standard deviation (SD) values of the processed wavelet detail coefficients were calculated. On the basis of the scattering theory, the resulting SD value, which represents the scattering intensity of the signal, is an indication of the fouling level. The laboratory results clearly show that the SD value decreases as the fouling level increases. The effectiveness of the proposed approach was then validated by field data from the Orin subdivision in Wyoming. By a comparison of in situ ground truth data and GPR measurements, the wavelet transform was proved to be an effective approach to quantifying ballast-fouling conditions. This approach allows for fouling assessment without the presence of a clear interface between clean and fouled ballast and reduces user dependency. In addition, the new approach is capable of processing GPR data automatically and continuously and provides the entire fouling profile along the tracks.

Railroad facilities play an important role in the transportation system (1). Every year large quantities of goods and coal and a great many people are transported along railroad tracks. To ensure the safety and serviceability of railroad facilities, billions of dollars are spent every year in the United States on railroad track maintenance. Railroad ballast, one of the main components of the track structure, is responsible for resisting vertical, lateral, and longitudinal forces; supporting railroad ties; reducing stress from the tie-bearing area; providing resiliency and energy absorption for the track; absorbing noise; providing immediate drainage; and alleviating the formation of frost (2). However, over time, ballast is gradually fouled by fine materials,

which usually come from ballast breakdown and infiltration of small particles. The phenomenon of the fine materials filling the air voids in the ballast is called fouling. When the fouling reaches a specific level, the structure integrity can be jeopardized and the drainage capacity can be undermined. Fouling can also cause structure instability or even train derailments. To optimize railroad maintenance and ensure the safety of railroad transportation, it is crucial to have a reliable technology to rapidly and accurately assess ballast-fouling condition.

Traditional methods for assessing ballast-fouling conditions include visual observation and sample drilling at intervals along the railroad track. However, these methods have certain disadvantages. Through surface visual survey, information underneath the track is usually missed. Although drilling samples provides reliable information about the ballast, drilling samples is time-consuming and can provide information only at certain discrete locations. To overcome these disadvantages, ground-penetrating radar (GPR) has been applied for detection of railroad ballast fouling. As a nondestructive test technique, GPR has shown its potential in assessing railroad ballast conditions effectively, rapidly, and continuously (2).

In regard to GPR's applications for railroads, various research studies have been conducted on GPR equipment selection, antennae setup, data management, and data processing and interpretation. Jack and Jackson used 450-MHz and 900-MHz GPR antennae to obtain the image attributes of the ballast and subgrade (3). The thickness of the ballast layer was estimated by assuming a fixed electromagnetic wave velocity of  $5.1 \times 10^8$  in./s ( $1.3 \times 10^8$  m/s). Results showed that significant horizons can be imaged, the changes along the horizons can be correlated with quality or structural variations, and the consistency of the depth of clean ballast could be monitored. Narayanan et al. compared GPR data collected from antennae with different frequencies (4). The 400-MHz antennae showed a better ability to estimate the depth of substructure layers than did the 100-MHz antennae. The 100-MHz antennae were better than the 400-MHz antennae for detecting the water pockets at deeper depths.

Olhoeft and Selig (5) and Sussmann (6) used 1-GHz air horn antennae to assess the rail track condition. They found that 1-GHz antennae were capable of obtaining information from the track substructure including thickness of ballast and subballast layers, variations in layer thickness along the track, and presence of water pockets trapped in the ballast and soft subgrade due to the high water content. Clark et al. (7) and Silvest et al. (8) used frequency spectrum and frequency sum techniques to assess ballast conditions under various fouling conditions.

Roberts et al. (9) and Al-Qadi et al. (10) used 2-GHz air-coupled antennae to assess the fouling condition on the basis of the scattering

---

Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, 205 North Mathews Avenue, MC-250, Urbana, IL 61801. Corresponding author: P. Shangguan, shanggu1@illinois.edu.

*Transportation Research Record: Journal of the Transportation Research Board*, No. 2289, Transportation Research Board of the National Academies, Washington, D.C., 2012, pp. 95–102.  
DOI: 10.3141/2289-13

response of EM wave in ballast. The scattering analysis approach showed potential to distinguish fouled ballast from clean ballast. To overcome the limited penetration depth problem of 2-GHz antennae, lower-frequency antennae such as 500-MHz horn antennae can also be used to obtain comprehensive information for the ballast, subballast, and subgrade (2). Al-Qadi et al. (11), Xie et al. (2), and Leng and Al-Qadi (12) used a time–frequency approach to interpret the GPR data. They found that this approach can characterize the signal in time and frequency domains simultaneously and quantify the fouling and moisture content.

In summary, GPR is a great tool for assessing the railroad ballast condition. However, the processing and interpreting of GPR data are challenging. Currently, there are three primary approaches for data interpretation: traditional approach, scattering approach, and time–frequency approach. The traditional approach interprets the data in the time domain. The thickness of the clean ballast can be determined if a clear interface between clean and fouled ballast is observed in the GPR image. The scattering approach interprets GPR data in the time domain and provides fouling depth on the basis of the difference in scattering intensity (10). The time–frequency approach applies short-time Fourier transform (STFT) to track the frequency change with time (2). When fouling is present, energy drops at the fouling location in the STFT image so the fouling depth can be obtained. These three methods can be combined for data interpretation, but the methods have certain limitations.

It is not unusual for the gradation of fouled ballast to change gradually, which results in no clear interface between clean and fouled ballast. Therefore, it is very difficult to obtain fouling depth with the traditional approach. When the scattering approach is used, fouling depth is determined at the location where scattering intensity changes. To see the change in scattering intensity, users need to choose the appropriate color map and display parameters of the GPR image. Yet different users may choose different color maps and parameters and thus interpret the information differently. Therefore, the result could be user dependent, especially when scattering is not clear enough. Although the time–frequency approach, which uses the STFT method, can provide more accurate fouling information, it can process only one scan per time (2). The time–frequency approach is suitable for assessing the fouling condition at specific locations, but it can be challenging to analyze the GPR data for an entire railroad network with this technique.

The objective of this study is to develop a new data processing technique that overcomes the aforementioned limitations. The wavelet technique, which is one of the most efficient data analysis techniques, is proposed for GPR data interpretation. Controlled laboratory tests were conducted at the Illinois Center for Transportation. Wavelet transform was then applied to the collected data to extract useful information that can indicate fouling conditions. Finally, the proposed wavelet technique was validated with field GPR data collected from the Orin subdivision in Wyoming.

## BACKGROUND

### GPR Principles

GPR has a transmitter antenna and a receiver antenna. The transmitter antenna sends out electromagnetic waves into the ground, and the receiver antenna collects the reflected signal scattered back from the interfaces and inhomogeneities having different electromagnetic properties within the materials. Figure 1 shows the electromagnetic wave propagation paths in railroad ballast. Path  $S_1$  represents the

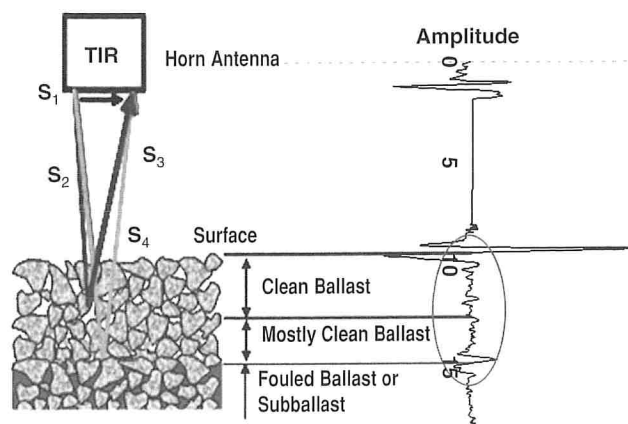


FIGURE 1 Typical GPR signal from ballast (2).

electromagnetic wave propagating directly from transmitter to receiver. This part of the signal is usually considered as noise and can be easily removed. Path  $S_2$  is the wave reflected by the surface of the ballast. Path  $S_3$  represents the electromagnetic wave scattered by the ballast–air–fouling interfaces. This part of the signal (the signal in the circle on the right of Figure 1) is propagating within the ballast and thus contains most of the fouling information. Path  $S_4$  represents the interface between clean ballast and fouled ballast or subballast. Path  $S_4$  may not be observed if the interface is not clear.

### Scattering Response

In clean ballast, the aggregate occupies about 70% of the volume and the air voids occupy the remaining 30%. Therefore, the aggregate can be considered as a transmission medium and the air voids as the scatter objects. Depending on the relationship between wavelength and the size of the scatterer, the electromagnetic wave scattering can be divided into three different domains. If the sizes of the scatterers are much larger or smaller than the wavelength, the scattering is grouped into the geometric scattering domain or the Rayleigh scattering domain. When the scatterers are on the scale of the wavelength, the scattering is in the Mie scattering domain (13). The value of the normalized dimension  $D^N$  can be used to compare the sizes of scatterers and the wavelength:

$$D^N = \frac{2\pi a}{\lambda} \quad (1)$$

$$\lambda = \frac{c}{f \cdot \sqrt{\epsilon_r}} \quad (2)$$

where

- $D^N$  = normalized dimension of the scatterer,
- $a$  = air void dimension,
- $\lambda$  = wavelength in ballast,
- $c$  = speed of light in free space ( $1.2 \times 10^{10}$  in./s =  $3 \times 10^8$  m/s),
- $f$  = GPR antenna dominant frequency, and
- $\epsilon_r$  = dielectric constant of the medium.

When the scatterer's size is comparable with the incident wavelength, the field diffraction in the shadow region and re-reflections between

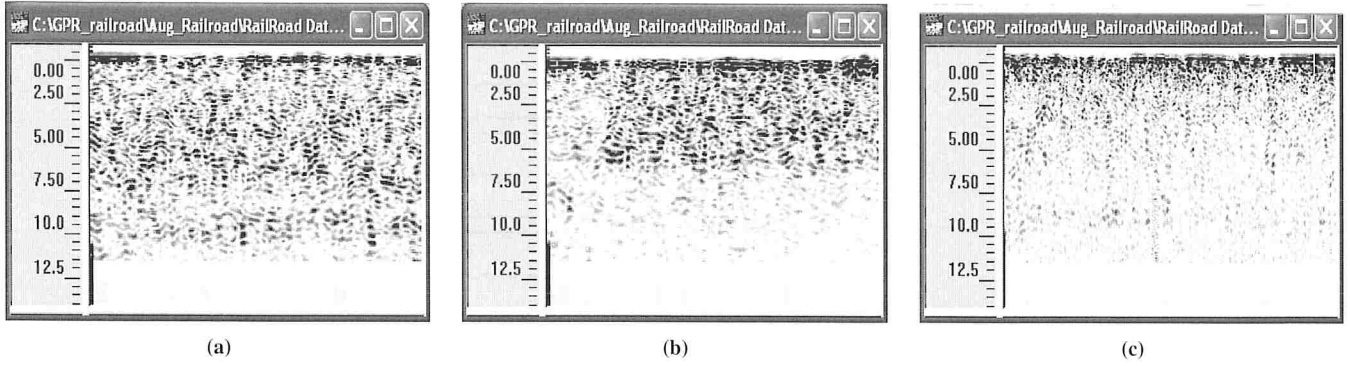


FIGURE 2 GPR images on ballast exhibiting various fouling conditions: (a) clean ballast, (b) moderately fouled ballast, and (c) fouled ballast (10).

local scatterers are significant (14). In clean ballast, the size of an air void varies from 11 to 29 mm. The related normalized dimension of an air void in clean ballast with a 2-GHz antenna is in the 0.5–1.2 range and in the 0.25–0.6 range when a 1-GHz antenna is used. So when a 2-GHz horn antenna is used, scattering is the dominant response in clean ballast, as shown in Figure 2a. With fouling materials filling the air voids, the normalized dimensions of air voids decrease. Thus the scattering response becomes weaker, as shown in Figure 2, b and c. The intensity of scattering in the ballast can be used to distinguish fouled ballast from the clean ballast.

## Wavelet Transform

In signal processing Fourier transform is used to show the frequency spectrum of the signal. But while the signal is transformed from the time domain to the frequency domain, the time information is lost. STFT can keep the data information in both the time and the frequency domain (12). However, STFT uses a window function that has a preset window length. Once this window function is chosen, the resolutions are fixed and kept the same at all frequencies and times. Wavelet transform is a multiresolution signal processing technique with adjustable window length. The wavelet technique has been successfully used in signal de-noising, data compression, and image processing (15, 16).

Continuous wavelet transform (CWT) decomposes a signal into a family of functions that offer good time and frequency localization. A mother wavelet  $\Psi(t)$  is used to carry out the wavelet decomposition. Typical mother wavelets can be found in the literature (15, 16). Wavelet function is defined in the following equations (17, 18):

$$\Psi_{a,\tau}(t) = \frac{1}{\sqrt{a}} \Psi\left(\frac{t-\tau}{a}\right) \quad a \in \mathbb{R}^*, \tau \in \mathbb{R} \quad (3)$$

$$C_f(a, \tau) = \int_{\mathbb{R}} f(t) \Psi_{a,\tau}^*(t) dt \quad (4)$$

where

- $t$  = time,
- $\Psi_{a,\tau}(t)$  = constructed by dilation and translation of mother wavelet  $\Psi(t)$ ,
- $a$  = parameter of dilation (or scaling),
- $\tau$  = parameter of translation (or shifting),

$C_f(a, \tau)$  = CWT of function  $f(t)$ , and  
superscript \* = complex conjugation.

Because CWT uses continuous scaling and shifting factors  $a$  and  $\tau$ , respectively, which produce an infinite number of wavelet coefficients, CWT is not computationally efficient. Therefore, discrete wavelet transform (DWT) is used in this study. Discrete wavelets are not continuously scalable and translatable. They can be scaled and shifted only in discrete steps. DWT is achieved by assigning scaling and shifting factors  $a$  and  $\tau$ , respectively, with the power of two values (19):

$$a = 2^j \quad (5)$$

$$\tau = k \cdot a = k \cdot 2^j \quad (6)$$

where  $j$  is the level at which the discrete wavelet analysis is performed and  $k$  is an integer parameter in DWT. By applying DWT on signal  $f(k)$ , two parts of the coefficients can be obtained: approximation coefficients and detail coefficients. The approximation coefficients are the low-frequency components, and the detail coefficients are the high-frequency components of signal  $f(k)$  (19). These two coefficients are expressed in the following equations:

$$A_j = \sum_{n=0}^{\infty} f(n) \phi_{jk}(n) = \sum_{n=0}^{\infty} f(n) \frac{1}{\sqrt{2^j}} \phi\left(\frac{n-k2^j}{2^j}\right) \quad (7)$$

$$D_j = \sum_{n=0}^{\infty} f(n) \frac{1}{\sqrt{2^j}} \Psi\left(\frac{n-k2^j}{2^j}\right) \quad (8)$$

where  $A_j$  and  $D_j$  are the approximation coefficient and detail coefficient, respectively, at level  $j$  and  $\phi_{jk}(n)$  is the scaling function associated with the wavelet function  $\Psi_{jk}(n)$ . As the decomposition level  $j$  increases, a hierarchical set of approximations and details can be obtained. This procedure is called multiresolution analysis (20, 21).

Figure 3 shows the flowchart of wavelet decomposition and wavelet coefficients of a typical GPR signal. In Figure 3a, the original signal, denoted by  $s$ , is decomposed into Level 1 approximation  $a_1$  and detail  $d_1$ , where  $a_1$  and  $d_1$  are calculated by Equations 7 and 8. Similarly,  $a_1$  is decomposed into Level 2 approximation  $a_2$  and detail  $d_2$ . After five levels of decomposition, the original signal  $s$  is decomposed into five details and one approximation and can be reconstructed with  $s = a_5 + d_5 + d_4 + d_3 + d_2 + d_1$ . In Figure 3b, the

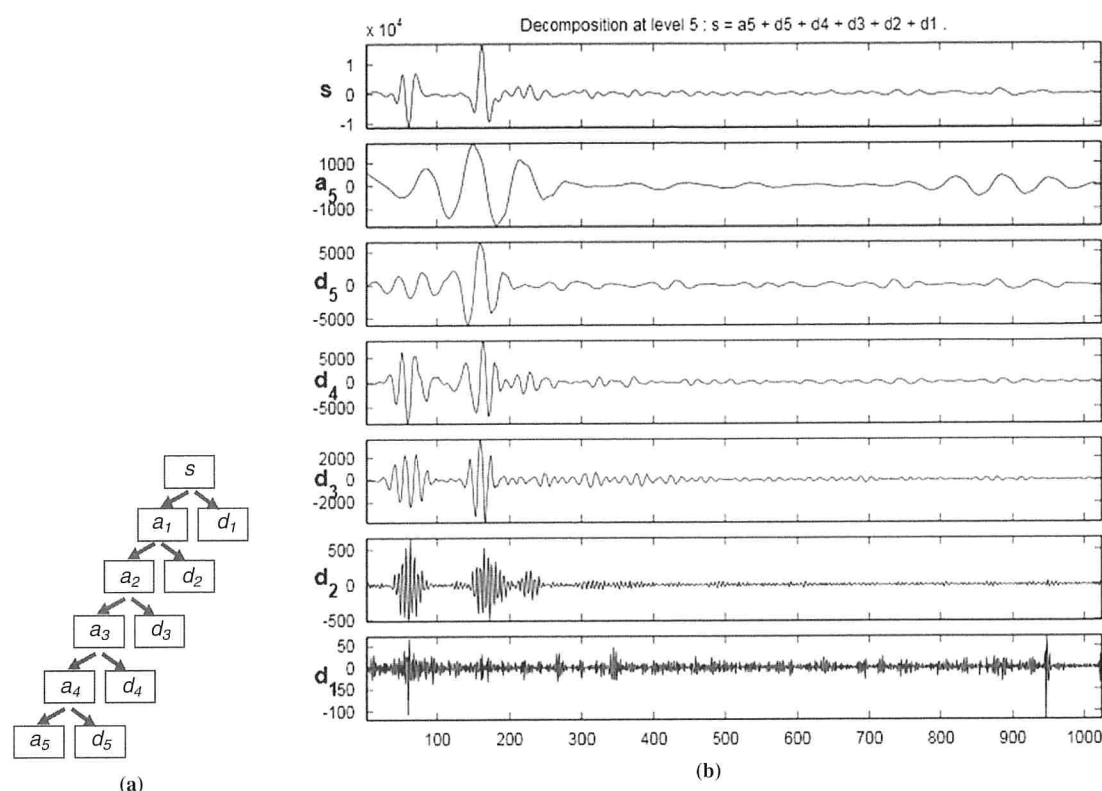


FIGURE 3 Wavelet decomposition: (a) five levels of wavelet decomposition of signal  $s$  flowchart and (b) five levels of wavelet decomposition of a typical GPR signal with Daubechies Order 5 (db5) mother wavelet.

GPR signal collected from a laboratory test is decomposed into five levels. It is easy to find that detail  $d_1$  is primarily high-frequency components with small magnitude (varies between  $-100$  and  $50$ ) and detail  $d_5$  is primarily low-frequency components with large magnitude (varies between  $-5,000$  and  $5,000$ ). Figure 3b will be discussed in more detail later.

## LABORATORY TESTS AND WAVELET TECHNIQUE INVESTIGATION

To investigate how well the wavelet technique interprets GPR data to assess the ballast condition, laboratory tests were conducted.

### Test Procedure

As shown in Figure 4a, a 5-ft  $\times$  5-ft  $\times$  4-ft (1.5-m  $\times$  1.5-m  $\times$  1.2-m) wooden box was built as a ballast container. The frame and all the screws are nonmetal materials so the container generates no noise. The box was filled with 40 in. (1,016 mm) of clean limestone ballast initially. A 2-GHz air-coupled GPR antenna and an SIR-20 system manufactured by Geophysical Survey Systems, Inc., were used to collect the data. A previous study showed that the limestone ballast was uniformly graded with an aggregate size of 2.5 in. (63.5 mm), the ballast air void content after compaction was 37.8%, and the dielectric constant of the clean ballast was 3.9 (12). Dry clay was spread evenly into the ballast at four levels of the air void volume, 10%, 20%, 30%, and 40%, as shown in Figure 4b. GPR data were collected at each fouling level.

## Wavelet Analysis

### Mother Wavelet Selection

The first step of wavelet transform is to choose a proper mother wavelet. Figure 5 shows several commonly used mother wavelets, including Haar, Biorthogonal Order 1 for reconstruction and Order 5 for decomposition (bior 1.5), and Daubechies Order 5 (db5). The wavelet coefficients actually describe how well the scaled and shifted mother wavelet matches the original signal. Thus, it is better to choose a mother wavelet that is as similar to the incident GPR wave as possible (20).

Comparing the three mother wavelets in Figure 5,  $a$ – $c$ , and the incident GPR wave in Figure 5d, db5 is similar to the incident wave and was chosen as the mother wavelet for GPR data analysis.

### Wavelet Coefficients Selection

After the mother wavelet is selected, the signal is decomposed into approximation and detail coefficients. However, not all the wavelet coefficients are useful. Some of them are noises. Selecting the appropriate wavelet coefficients is critical to signal de-noising and data interpretation.

By using db5 mother wavelet for a five-level wavelet decomposition, five approximation coefficients,  $a_1$ ,  $a_2$ ,  $a_3$ ,  $a_4$ ,  $a_5$ , and five detail coefficients,  $d_1$ ,  $d_2$ ,  $d_3$ ,  $d_4$ ,  $d_5$ , are obtained. Each level of the detail has a pseudofrequency that is determined by the sample frequency of the signal, mother wavelet type, and decomposition level. For the GPR settings in this test, the total time range for one scan is



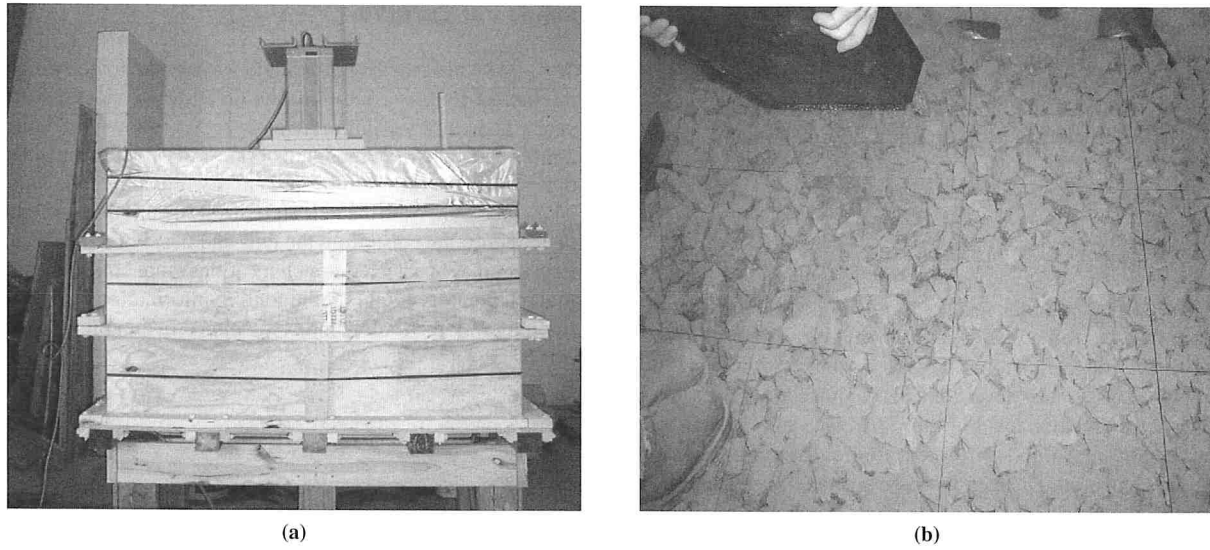


FIGURE 4 Laboratory experiment to collect GPR data on ballast with controlled fouling levels: (a) ballast box and 2-GHz air-coupled antenna placed on top of ballast and (b) clay spread evenly into ballast.

30 ns and each scan has 1,024 samples. So the sample frequency is  $1,024 - (30 \text{ ns}) = 34.1 \text{ GHz}$ . The corresponding pseudofrequency for each level of the detail coefficient is 11.4 GHz for Level 1, 5.7 GHz for Level 2, 2.8 GHz for Level 3, 1.4 GHz for Level 4, and 0.7 GHz for Level 5. As seen in Figure 3b, the frequency from  $d_1$  to  $d_5$  decreases.

Because the central frequency of the GPR antenna used in this test is 2 GHz, most of the EM wave energy is concentrated around 2 GHz. As shown in Figure 3b, Levels 1 and 2 details have high pseudofrequencies of 11.4 GHz and 5.7 GHz and thus are mainly noise. These two levels of detail coefficients should be removed. The other three levels of detail coefficients have the pseudofrequencies of 0.7,

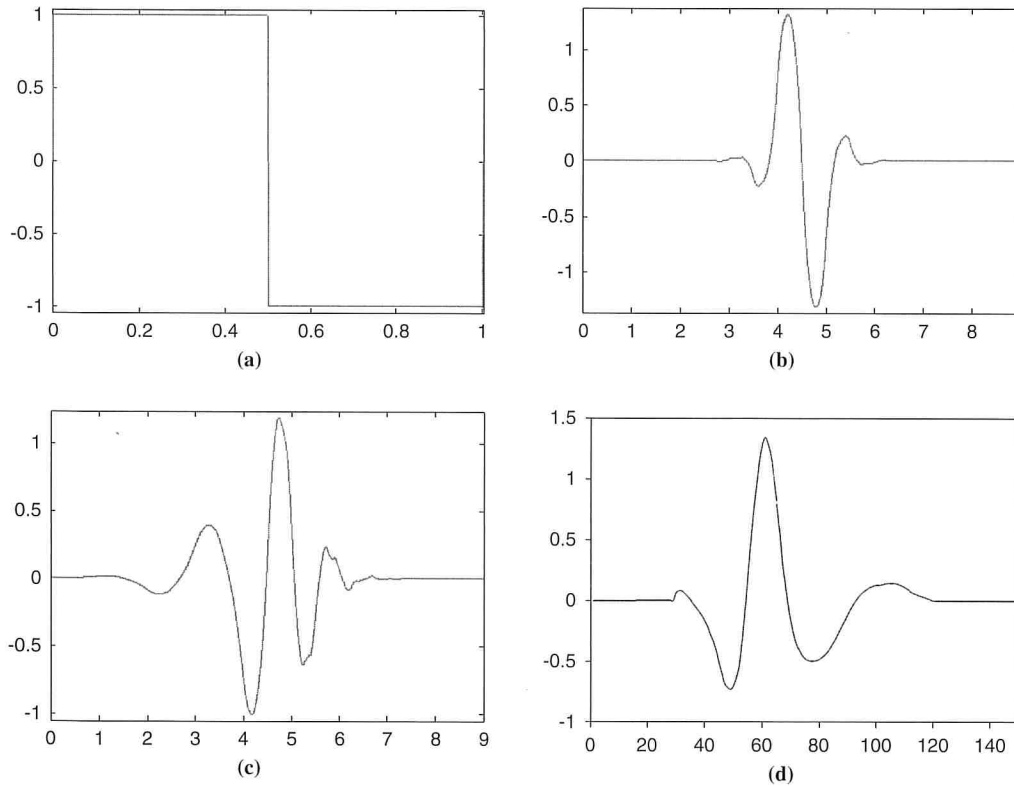


FIGURE 5 Mother wavelets and incident GPR wave: (a) Haar, (b) bior 1.5, (c) db5, and (d) incident GPR wave.



1.4, and 2.8 GHz, which are close to 2 GHz. As mentioned in the scattering theory, for the 2-GHz EM wave, the scattering response is dominant in clean ballast and the change of scattering intensity caused by the change of air void size is significant. Thus, these three wavelet detail coefficients contain fouling information and should be analyzed.

### Standard Deviation Calculation

The level of fluctuation of the received signal reflects the scattering intensity of the electromagnetic wave when it propagates through the ballast. The more the signal fluctuates, the more intense the scattering is. The standard deviation (SD) value can be calculated to evaluate the fluctuation level of the signal, and thus the scattering intensity for ballast under different fouling conditions can be obtained and compared. The data processing procedure is described as follows:

- For GPR data of 10% fouling level, apply discrete wavelet transform by using the db5 mother wavelet for five levels of decomposition on one scan;
- Combine Level 3, 4, and 5 wavelet detail coefficients, which have a pseudofrequency around 2 GHz; this step also de-noises the signal;
- Remove coupling pulse and surface reflection in the three detail signals because they do not reflect information of the fouling;
- Calculate SD values for the processed signal;
- Repeat the above steps for other scans and determine the average of the SD values for all scans; although other scans were also collected at the same location, this step can significantly reduce the randomness of the GPR ballast signal; and
- Repeat the above steps for GPR data of 20%, 30%, and 40% fouling levels. Compare the SD values with the fouling levels.

Figure 6 shows the resultant SD values. As the fouling level increases from 10% to 40%, the SD value decreases from 534 to 469. This result can be explained by the scattering theory. As the fouling materials fill the air voids in the ballast, the scattering intensity decreases and the SD value of the processed signal that represents the scattering intensity also decreases. Compared with the scattering approach and the time–frequency approach, which intend to obtain the specific fouling depth, the wavelet approach tends to evaluate the fouling level through the whole ballast layer. The laboratory GPR results show that the wavelet technique has great potential for estimating fouling levels.

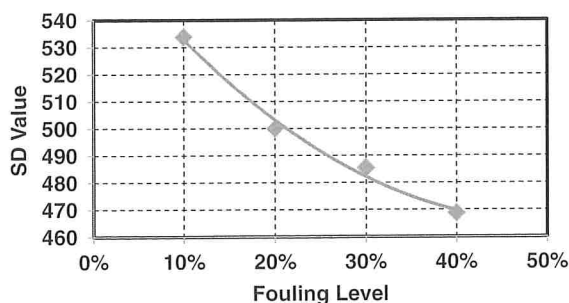


FIGURE 6 SD values for fouling levels of 10%, 20%, 30%, and 40%.

### FIELD VALIDATION

To validate the effectiveness of the wavelet technique for assessing the ballast-fouling condition, the GPR data collected at the Orin subdivision in Wyoming were analyzed. As shown in Figure 7, two 2-GHz antennae were mounted on a hi-rail vehicle suspended above the rail track to allow for rapid survey. Each antenna was placed 24 in. (600 mm) from the rails and 6 in. (150 mm) from the edge of the tie to reduce the noise from the rails and ties. For research purposes, field samples at selected trench locations were collected and sieved in the laboratory to obtain the fouling information. To describe the fouling level, the fouling index as defined in the following equation was used ( $I$ ):

$$F_I = P_4 + P_{200} \quad (9)$$

where

$F_I$  = fouling index,

$P_4$  = weight percentage of particles passing the 4.75-mm (No. 4) sieve, and

$P_{200}$  = percentage of fine particles passing the 0.075-mm (No. 200) sieve.

If  $F_I$  is smaller than 10%, the ballast is considered to be clean ballast. If  $F_I$  is between 10% and 20%, the ballast is moderately fouled. If  $F_I$  is larger than 20%, the ballast is seriously fouled.

The GPR data were processed by using the aforementioned procedure. The SD values and corresponding fouling indices are reported in Table 1. In the location column, the number after MP (milepost) indicates the location of the trench. The data are plotted in Figure 8. As expected, the field data show the same trend as the laboratory data. When the fouling index increases from 11.7% to 40.8%, the SD value decreases from 636 to 355. This validates the effectiveness of the wavelet technique and SD value to indicate fouling conditions.

By using the results in Figure 8, the following regression relation between  $F_I$  and the SD value can be obtained:

$$F_I = 3.043 \times 10^{-6} \times \text{SD}^2 - 3.924 \times 10^{-3} \times \text{SD} + 1.394 \quad (10)$$

where  $F_I$  is the fouling index and SD is the standard deviation value of the processed signal. The coefficient of determination  $R^2$  of this

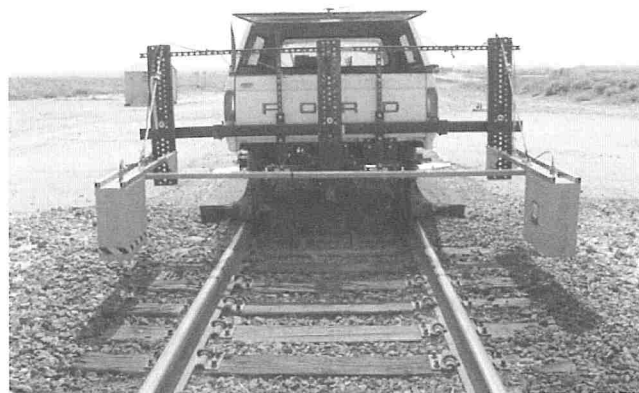


FIGURE 7 GPR antennae setup on railroad track.

TABLE 1 Fouling Indices and Corresponding SD Values at Different Locations

Location	Fouling Index (%)	SD Value
MP 35.900	15.8	573
MP 36.000	11.7	636
MP 37.000	40.8	355
MP 37.855	23.8	434
MP 42.000	21.6	502
MP 52.054	28.5	402

NOTE: MP = milepost.

equation is .95 and the standard error of estimation is 0.02, which indicate good correlation between the variables. One advantage of the wavelet approach is that it can process all the scans in the GPR data file automatically, continuously, and rapidly. The whole procedure of the wavelet approach and the resultant fouling profile are shown in Figure 9. As shown, the fouling profile along the distance can be generated. From Milepost 35.55 to Milepost 35.65,  $F_f$  is larger than 10% and the ballast is moderately fouled. From Milepost 35.65 to Milepost 35.82,  $F_f$  is smaller than 10%, which means the ballast is clean. The fouling profile provides helpful information for identifying sections that require maintenance and optimizing maintenance funds.

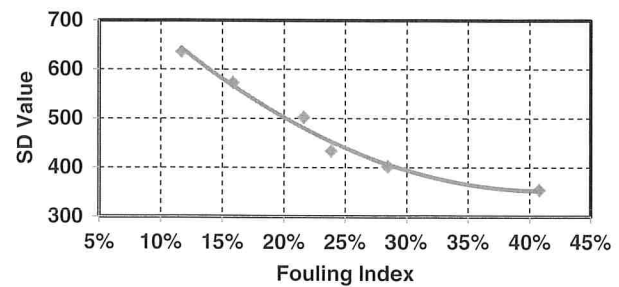


FIGURE 8 SD values for different fouling indices at trench locations.

## CONCLUSIONS

GPR is an effective and efficient tool that can be used for assessing railroad track condition rapidly and nondestructively. This study proposes a new approach, DWT, to process and interpret the GPR data to estimate ballast-fouling conditions. Controlled laboratory tests were conducted to examine the introduced approach. Wavelet transform was then applied to the signal to obtain the wavelet coefficients. SD values of the wavelet coefficients were calculated to quantify the scattering intensity. The SD value is capable of indicating the fouling level of the ballast. According to the results

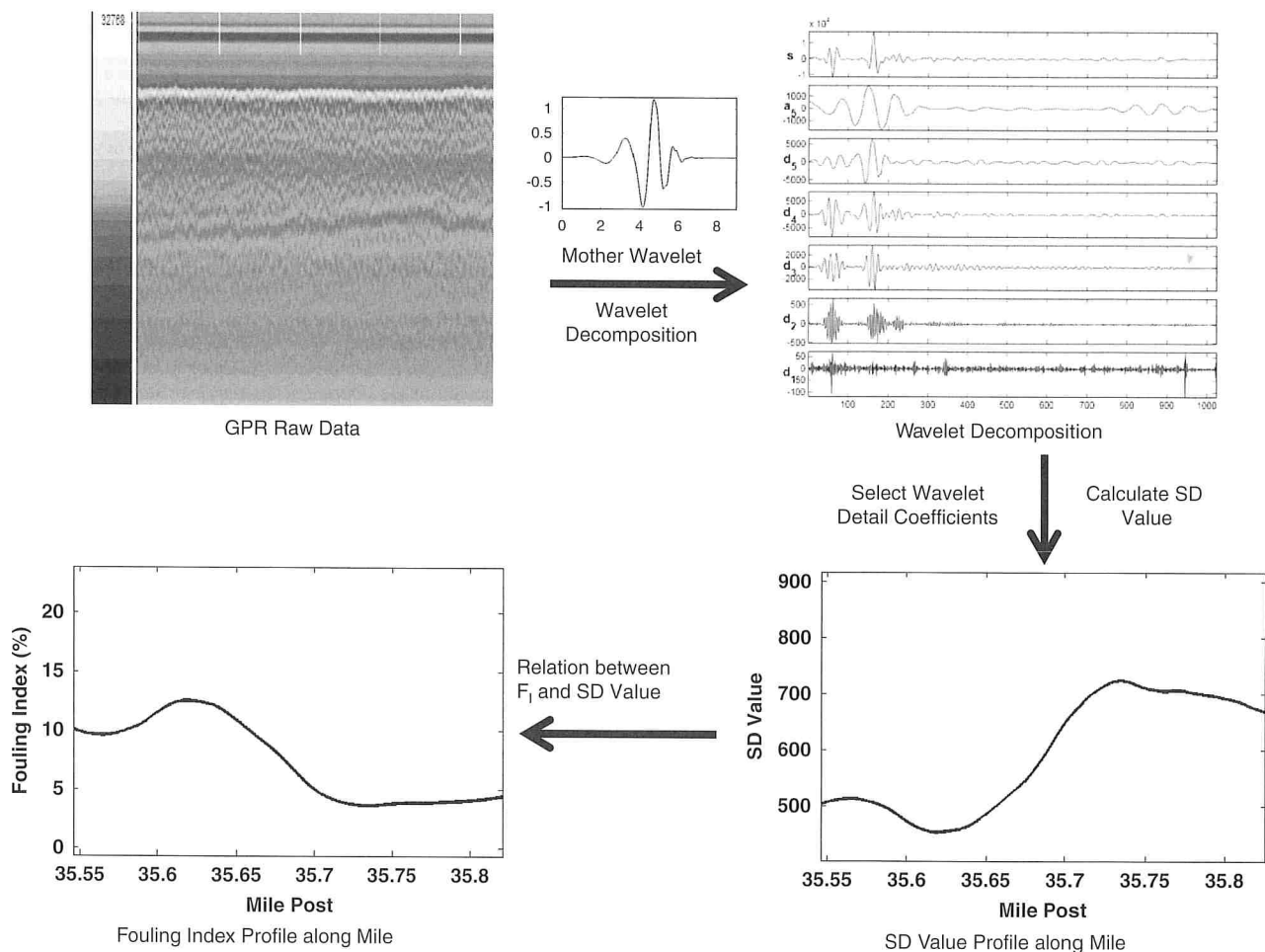


FIGURE 9 Procedure for fouling quantification by using wavelet approach.

from both laboratory and field GPR data, from clean ballast to fouled ballast, the SD value varied from 700 to 350. As the fouling level of the ballast increases, the SD value decreases. This decrease can be explained by the scattering theory. This approach has been validated by field GPR data and ground truth measurements. The new approach has the following advantages compared with other approaches:

- The new approach can estimate the fouling condition without clear interface between clean and fouled ballast. The traditional approach, which interprets the data in the time domain, is not able to provide the fouling location if the gradation of the ballast changes gradually with depth. The new approach calculates the scattering intensity in the whole ballast layer and can quantify the fouling without interface.

- The new approach can significantly reduce the user dependency of GPR data interpretation results. In the scattering approach, the GPR data are interpreted on the basis of scattering images in the time domain. Therefore, the fouling location is sometimes operator dependent. Although the STFT approach can provide more accurate images in the STFT domain, the fouling locations can still vary with different time window lengths. In the wavelet approach, by following the same data processing procedure, a single parameter, the SD value, can be obtained to judge the fouling levels. This parameter is operator independent.

- The new approach allows for group processing of GPR data and can generate a fouling profile at the network level for railroad tracks. Although the STFT approach assesses the ballast-fouling condition accurately, it can process only one scan at a time. The new approach processes all the scans in the GPR files quickly and automatically. An entire fouling profile for a defined distance can be obtained. However, further studies are needed to obtain more accurate relations between SD values and fouling indices for various ballast types and GPR data collection settings.

The moisture effect is not considered in the developed algorithm. The strong correlation between SD values and fouling indices could be affected by ballast type, type of fouling material, and moisture presence in ballast. More field data are needed to further validate the proposed algorithm under various conditions.

## ACKNOWLEDGMENTS

This research is sponsored by the University of Illinois at Urbana-Champaign (UIUC)–Association of American Railroads (AAR) Technology Scanning Research Program. The invaluable input of Roger Roberts, Erol Tutumluer, and Samer Lahouar is greatly appreciated. The authors also acknowledge the assistance of James Meister and the research project panel for their help and feedback during the study.

## REFERENCES

1. Selig, E. T., and J. M. Waters. *Track Geotechnology and Substructure Management*. Thomas Telford Ltd., London, 1994.
2. Xie, W., I. L. Al-Qadi, D. L. Jones, and R. L. Roberts. Development of a Time-Frequency Approach to Quantify Railroad Ballast Fouling Condition Using UWB GPR Data. Presented at 87th Annual Meeting of the Transportation Research Board, Washington, D.C., 2008.
3. Jack, R., and P. Jackson. Imaging Attributes of Railway Track Formation and Ballast Using Ground Probing Radar. *NDT & E International*, Vol. 32, 1999, pp. 457–462.
4. Narayanan, R. M., C. J. Kumke, and D. Li. Railroad Track Monitoring Using Ground Penetrating Radar: Simulation Study and Field Measurements. *Proceedings of SPIE Conference on Subsurface Sensors and Applications*, Vol. 3752, Denver, Colo., July 1999.
5. Olhoeft, G. R., and E. T. Selig. Ground Penetrating Radar Evaluation of Railroad Track Substructure Conditions. *Proc., 9th International Conference on Ground Penetrating Radar*, Santa Barbara, Calif., 2002, pp. 48–53.
6. Sussmann, T. R. *Application of Ground Penetrating Radar to Railway Track Substructure Maintenance Management*. PhD dissertation. University of Massachusetts, Amherst, 1999.
7. Clark, M. R., R. Gillespie, T. Kemp, D. M. McCann, and M. C. Forde. Electromagnetic Properties of Railway Ballast. *NDT & E International*, Vol. 34, No. 5, 2001, pp. 305–311.
8. Silvast, M., M. Levomaki, A. Nurmikolu, and J. Noukka. NDT Techniques in Railway Structure Analysis. Presented at World Congress on Railway Research Conference, Montreal, Quebec, Canada, 2006.
9. Roberts, R., I. L. Al-Qadi, E. Tutumluer, J. Boyle, and T. Sussmann. Advances in Railroad Ballast Evaluation Using 2-GHz Horn Antenna. Presented at 11th International Conference on Ground Penetrating Radar, Columbus, Ohio, 2006 (CD).
10. Al-Qadi, I. L., W. Xie, and R. Roberts. Scattering Analysis of Ground-Penetrating Radar Data to Quantify Railroad Ballast Contamination. *Journal of Nondestructive Testing and Evaluation*, Vol. 41, No. 6, 2008, pp. 441–447.
11. Al-Qadi, I. L., W. Xie, R. Roberts, and Z. Leng. Data Analysis Techniques for GPR Used for Assessing Railroad Ballast in High Radio-Frequency Environment. *Journal of Transportation Engineering*, Vol. 136, No. 4, 2010, pp. 392–399.
12. Leng, Z., and I. L. Al-Qadi. Railroad Ballast Evaluation Using Ground-Penetrating Radar: Laboratory Investigation and Field Validation. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 2159, Transportation Research Board of the National Academies, Washington, D.C., 2010, pp. 110–117.
13. Chuang, S. L. *Physics of Photonic Devices*. John Wiley & Sons, Inc., New York, 2009.
14. Astanin, L. Y., and A. A. Kostylev. *Ultrawideband Radar Measurements: Analysis and Processing*. Institute of Electrical Engineers, London, 1997.
15. Chui, C. K. *An Introduction to Wavelets*. Academic Press, New York, 1992.
16. Daubechies, I. *Ten Lectures on Wavelets*. CBMS-NSR Series in Applied Mathematics, SIAM, 1992.
17. Donoho, D. L., and I. M. Johnstone. Adapting to Unknown Smoothness via Wavelet Shrinkage. *Journal of the American Statistical Association*, Vol. 90, No. 432 1995, pp. 1200–1224.
18. Young, R. K. *Wavelet Theory and Its Applications*. Kluwer Academic Publishers, Boston, Mass., 1993.
19. Zhang, H., and T. R. Blackburn. A Novel Wavelet Transform Technique for On-line Partial Discharge Measurements, Part 1: WT De-noising Algorithm. *IEEE Transactions on Dielectrics and Electrical Insulation*, Vol. 14, No. 1, 2007, pp. 3–14.
20. Baili, J., S. Lahouar, M. Hergli, I. L. Al-Qadi, and K. Besbes. GPR Signal De-noising by Discrete Wavelet Transform. *NDT & E International*, Vol. 42, No. 8, 2009, pp. 696–703.
21. Mallat, S. G. A Theory for Multiresolution Signal Decomposition: The Wavelet Representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 11, No. 7, 1989, pp. 674–693.

*The contents of this paper reflect the view of the authors, who are responsible for the facts and the accuracy of the data. The contents do not necessarily reflect the official views or policies of the Association of American Railroads or the UIUC-AAR Program. This paper does not constitute a standard, specification, or regulation.*

*The Railway Maintenance Committee peer-reviewed this paper.*

# Stochastic Rail Wear Model for Railroad Tracks

Seosamh B. Costello, Anuradha S. Premathilaka, and Roger C. M. Dunn

**This paper describes the development of a stochastic rail wear model. Once validated, the model can be used to assist in the strategic assessment of railroad track funding needs. The algorithms used for simulating rail wear use Markov processes, and the resulting transition probability matrix defines rail wear progression, as opposed to the more familiar regression-type model popular with engineers. The New Zealand railroad track database contained 10 years of rail wear data from which to develop and validate the model. The transition probability matrices for use in the model were developed with the first 5 years of the historical rail wear data, with the remaining 5 years set aside to validate the model. The development of the transition probability matrices is reported in the paper, together with the development of the initial condition distributions ready for use in validation of the model.**

Recent changes in the ownership and management of New Zealand's rail infrastructure have enabled a longer-term view to be taken in the management of its asset. Although the railroad infrastructure was privatized between 1994 and 2004, it has now reverted back to public ownership and is managed by the New Zealand Railways Corporation (NZRC). NZRC is required to produce long-term strategic plans for the management of the rail track infrastructure, to demonstrate proper management of the public asset. Long-term strategic planning requires knowledge of not only the current condition of the asset, but also of its future condition determined from performance modeling.

NZRC's track database houses the inventory and current condition information of the track components; however, the predictive capability required to contribute to a long-term strategic plan was heretofore not available. This prompted a research study to develop deterioration models for New Zealand's rail track infrastructure. To date the study has focused on rail wear, rail being one of the most expensive track renewal items in New Zealand. It is recognized that rail defects also prompt renewals; however, this paper addresses only the mechanism of rail wear.

Deterministic rail wear models have been developed by Premathilaka et al.; however, the nature and variability of the data set suggest that a stochastic model would be more appropriate than a deterministic model in the prediction of rail wear at the network level (1). In particular, rail wear would be better represented at the network level by a stochastic process such as the Markov model,

similar to that described in Costello et al. for road pavements (2). Although Zobory explains how the principles of Markov theory, semi-Markovian stochastic models in particular, have been considered for profile wear prediction under specified operating conditions, a comprehensive literature review found that the Markov technique has not heretofore been applied to modeling network-level distributions of rail wear (3).

## STOCHASTIC NATURE OF RAIL WEAR

The available literature also suggests that rail wear is stochastic in nature, the premise being that the wear process of rails typically belongs to the class of stochastic phenomena because of the inherent uncertain character of the wheel-rail contact events realized in the course of standard railway operation (4, 5). Some of the uncertainty can be explained as the result of variations in rail lubrication effectiveness, train speed, track geometry, and track structure, as discussed below.

Rail lubrication, for example, is affected by many stochastic parameters that inhibit its effectiveness, such as the frequency of trains, frequency of lubrication, and the amount, type, and method of lubrication. Thelen and Lovette imply that the effectiveness of lubrication is directly influenced by many factors, including wheel and rail contour, rail geometry, dynamic characteristics of the track, surface conditions of the wheel and rail, viscosity and lubricity of the grease, operating temperature of the wheel and rail, and environmental factors such as temperature and precipitation, and indirectly influenced by the operating characteristics of the lubricators, train action, wheel slip, environmental contamination, and human factors (6). These variable influences result in an inability to have controlled and regular rail wear, which indicates that lubrication has a variable effect on rail wear (7).

Furthermore, Thelen and Lovette identified that lubrication of curved rail tracks, in particular, faces a number of difficulties in regard to its effectiveness (6). Laboratory tests have observed a considerable variation in the effectiveness of lubrication with distance from the applicator (8). Track lubrication in New Zealand is carried out by using high rail vehicles and trackside lubricators, and also by using various lubricator materials (9). Because of those factors, variation in the effectiveness of lubrication is inevitable.

The actual forces exerted on the track structure when a train traverses a curve depend very much on the speed at which the train negotiates the curve. Although each curve has a posted speed, each train will travel at a slightly different speed. The factors that influence this difference range from simple characteristics, such as driver characteristics, driver behavior, train characteristics, and weather conditions, to complex characteristics, as mentioned in Zobory and Szabo, such as wheel profiles, rail profiles, individual wheel loads, track stiffness, and track geometry (5).

---

S. B. Costello and R. C. M. Dunn, Department of Civil and Environmental Engineering, University of Auckland, Private Bag 92019, Auckland 1142, New Zealand. A. S. Premathilaka, Consultancy Services, W.D.M. Limited, North View, Staple Hill, Bristol BS16 4NX, United Kingdom. Corresponding author: S. B. Costello, s.costello@auckland.ac.nz.

*Transportation Research Record: Journal of the Transportation Research Board*, No. 2289, Transportation Research Board of the National Academies, Washington, D.C., 2012, pp. 103-110.  
DOI: 10.3141/2289-14



Track geometry is also strongly associated with the rate of rail wear. Track geometry maintenance activities are carried out when geometry characteristics exceed standard limits. However, it must be remembered that track geometry characteristics such as line, top, gauge, cant, and twist fluctuate within the allowable limits as a result of fluctuations in temperature and subgrade condition, among others. Detwiler and Nagurka suggest that track geometry possesses unintentional variations such as irregularities in alignment, gauge, and cross level, resulting in a stochastic input being used in rail vehicle dynamic simulation studies (10). From this result it can be inferred that track geometry is inherently stochastic.

Finally, track component performance is highly dependent on the magnitude and variation of subgrade stiffness, which itself is affected by many factors that have random properties (11). Oscarsson observed high variability in sleeper spacing, rail pad stiffness, ballast stiffness, and covibrating mass of the ballast subgrade, which resulted in their being modeled as random variables in that study (12).

## RAIL WEAR DATA

Rail wear data in New Zealand are collected manually by using a rail wear gauge. Typically, rail wear is monitored on curved tracks, where the rate of rail wear is fastest. On 50-kg/m rails, high-leg top wear (HT), low-leg side wear (LS), high-leg side wear (HS), and low-leg top wear (LT) are recorded. However, as HS and LT are dominant on curves, it was decided to exclude HT and LS from this study (9). High leg is the outside rail and the low leg is the inside rail at curved track sections. For 91- and 90-lb/yd rails, high leg and low leg readings are recorded as the total wear on the sides and the top. These readings are referred to as high leg total wear (HLT) and low leg total wear (LLT). Both were included in this study.

There are currently two types of rail wear gauges in use. For 50-kg/m rails the units of measurement on the gauge are millimeters. For all rails defined in imperial units, including 91- and 90-lb/yd rails, rail wear is measured by using a gauge with a point scale. Because the imperial gauge is designed for a number of rail types, a correction is applied to the reading. When measured by this gauge, brand-new 91- and 90-lb/yd rails record a total of 6 points; therefore all wear readings taken from this gauge have values of 6 or more. For statistical analysis this value needed to be taken into consideration. Both measurements of rail wear relate to the percentage of total head loss from the rail head.

All rail wear measurements are recorded in the track database, including historical readings. Each rail wear record contains a location reference, inspection date, and details of the type of rail. By cross-referencing the location reference with other files in the database, information such as installation date, annual tonnage, curve radius, posted speed, and cant, among other data, can all be linked to the record by using database queries.

The data set used in the analysis consisted of Class A principal lines, with an annual tonnage range between 1 and 5 million gross tonnes, on curves with less than an 800-m radius. Because of the filtering for radii less than 800 m, the data set contained curve posted speeds between 35 and 95 km/h and cant on the curves of either 70 mm (for curves less than or equal to 460 m) or 60 mm (for curves between 460 and 800 m).

At the time the research was carried out the track database contained 10 years of historical data from which to develop the model. However, data used in model development cannot be reused in validation of the model. Consequently, the available data were divided

into two equal time periods, with the transition probabilities determined from the first 5 years of data, and the model validated by using data from the remaining 5 years.

## STOCHASTIC MODELING

### Probabilistic Modeling

Probabilistic models predict the condition as the probability of occurrence of a range of possible outcomes. In situations in which the expected outcome is of a probabilistic nature, or in which there are insufficient data to develop reliable deterministic models, probabilistic methods offer a solution. There are two types of probability functions, either a continuous probability function or a Markovian process (13). The Markov process describes a changing system in which the future behavior is influenced by the present condition (14). A changing system could, for example, be the changing condition of an infrastructure asset. The main probabilistic method that has been used for modeling the management of infrastructure assets has been the Markov process (15).

The theory of Markov chains has been used extensively in modeling deterioration of infrastructure systems that are probabilistic in their deterioration pattern at the network level, such as in electricity networks, water networks, bridge management systems, and pavement management systems. In the highway sector, for example, probabilistic models based on the Markov technique are widely cited in the literature (2, 16–25).

### Markov Process

The Markov process is named after A. A. Markov, who introduced the concept in 1907; and later the “denumerable case,” commonly referred to as Markov chains, was launched by Kolmogorov in 1936 (26). A brief outline of the Markov process follows.

A collection of random variables with a probability distribution is called a “stochastic” process. A Markov chain is a special type of stochastic process, governed by the following three restrictions:

- The process is discrete in time—the dependent variable is random and can take only one of a finite or countable number of values in time.
- The process has a finite state space—there are countable possibilities of states or condition types of the dependent variable.
- The process satisfies the “Markov property”—the Markov property is attributed to a process in which its future state depends on the present state but not its past states (27).

A stochastic process that satisfies the above restrictions is called a discrete-time Markov chain. The phrase “discrete-time” comes from the first restriction, the word “chain” comes from the second restriction, and the word “Markov” comes from the third restriction. Furthermore, the process is said to be stationary (or homogeneous) in time, if the probability of moving from one state to another is independent of the time at which the step is being made (27). In this process, a probability is assigned for a system to move from one state to another.

In simple terms, the Markov process is executed by using two matrices, an initial condition vector (ICV) that describes the initial state of the network, and a transition probability matrix (TPM) that



describes the movements of the various components of the network to the next state in a given duty cycle. One cycle of deterioration is simulated by multiplying the ICV by the TPM. The repeated multiplication by the TPM is referred to as the Markov chain.

### Applicability of Markov Chains to Modeling Rail Wear

The following explains how the three restrictions in discrete-time Markov chains can be satisfied to apply this technique for modeling the rail wear process at the network level:

- The rail wear process occurs continuously throughout the year; therefore it can be considered continuous in time. However, rail wear readings are taken at yearly intervals, and therefore this process in turn is rendered discrete in time.
- The number of possible outcomes (state space) for rail wear is infinite between zero and the maximum possible rail wear at any given time. To achieve a finite state space, all possible outcomes are grouped into a finite number of bands of rail wear. This satisfies the restriction that the process has a finite space.
- The Markov property requires that if a particular rail is worn to a certain level, the future wear can depend only on its current condition but not on its past condition. It is assumed that the Markov property is satisfied for the wear of rails.

Stationary Markov chains (i.e., homogeneous in time) will be applied in developing the rail wear model. Stationary Markov chains consider that the rail will always wear according to the probabilities of a single TPM.

### Components of a Markov Chain

The two basic components of a Markov chain are the ICV and the TPM. The ICV represents the initial condition distribution of the rail network, in which the overall network rail length is distributed as percentages into a number of condition bands, depending on the amount of wear they have. This distribution can be determined relatively easily from the data on the current condition of the network contained in the track database. The general form of the ICV is shown in Equation 1 below:

$$a_0 = [\alpha_1 \quad \alpha_2 \quad \alpha_3 \quad \dots \quad \alpha_n] \quad (1)$$

There are two conditions that this vector ( $a_0$ ) needs to satisfy:

- The sum of all  $\alpha_i$  must be equal to one and
- All  $\alpha_i$  values must be positive.

The TPM is the controlling feature of the Markov chain and can be difficult to determine with confidence. A generic TPM is shown in Equation 2 in matrix notation. A TPM defines the probabilities that sections in a certain condition band move to another condition band in one duty cycle. A duty cycle is usually defined as a period of environmental exposure and traffic loading. The value  $p_{ij}$  is the probability of the proportion of the network currently in condition band  $i$  moving to condition band  $j$  after one duty cycle. In New Zealand, rail maintenance and rail renewal budgets are determined annually, and rail engineering inspections that monitor and record

rail wear are carried out annually; therefore, a year has been selected as the duty cycle for modeling purposes.

$$P = \begin{bmatrix} p_{11} & p_{12} & \dots & p_{1n} \\ p_{21} & p_{22} & \dots & p_{2n} \\ \vdots & & & \vdots \\ p_{n1} & p_{n2} & \dots & p_{nn} \end{bmatrix} \quad (2)$$

Similar to  $a_0$ , there are two conditions that TPMs need to satisfy:

- The sum of all entries in each row must be equal to one and
- All  $p_{ij}$  must be positive values.

The amount of rail wear can only increase with time; that is, rail wear does not reduce as rails deteriorate, unless they are replaced. As it is the rail deterioration that is being modeled, a reduction of the amount of rail wear cannot be allowed in the model. For that reason, all  $p_{ij}$  need to be set to zero where  $i > j$ . Furthermore,  $p_{nn}$ , the bottom rightmost value in the TPM, will always be equal to one because the rail has worn to its worst level and even if it wears further it will still be included in the same band. A TPM for an infrastructure asset takes the general form shown in Equation 3, in which all the  $p_{ij}$  have any value between zero and one, subject to the above restrictions.

$$P = \begin{bmatrix} p_{11} & p_{12} & p_{13} & \dots & p_{1n} \\ 0 & p_{22} & p_{23} & \dots & p_{2n} \\ 0 & 0 & p_{33} & \dots & p_{3n} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix} \quad (3)$$

Once the ICV ( $a_0$ ) and the TPM ( $P$ ) are known, the Markov process can be instigated. The distribution of condition after one duty cycle ( $a_1$ ) can be determined with Equation 4.

$$a_1 = [a_0 \times P] \quad (4)$$

Similarly, the distribution of condition after two duty cycles can be determined by multiplying the resulting vector ( $a_1$ ) by the same TPM ( $P$ ), because the process is a stationary Markov chain. Consequently, the distribution of rail wear after  $t$  years can be determined with Equation 5.

$$a_t = [a_0 \times P^t] \quad (5)$$

The process above can be considered the basic Markov modeling procedure, and it can be encapsulated relatively easily in a simple spreadsheet or a track asset management system.

## MODEL DEVELOPMENT

### Homogeneity

Performance predictions using Markov models are more accurate when homogeneous subnetworks are defined (19, 21, 28–31). However, this definition needs to be balanced by the amount of data

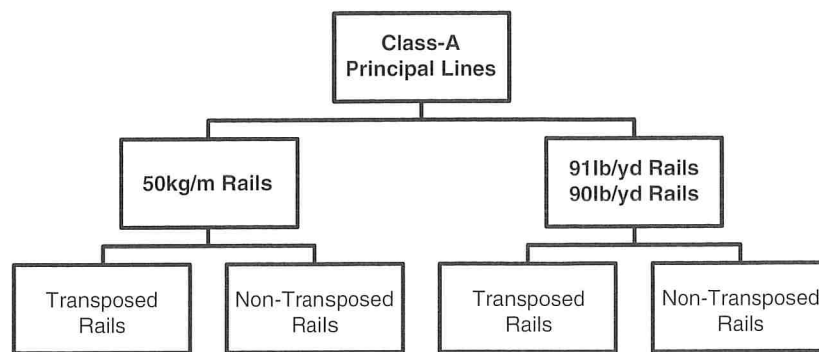


FIGURE 1 Subnetworks under consideration.

available to develop the model, in that too many subnetworks result in insufficient data from which to develop probabilities.

The focus of network-level rail wear modeling in this research is only on Class A principal lines in New Zealand, a homogeneous subnetwork in itself. It was also considered unwise to amalgamate 50-kg/m rails and 91- and 90-lb/yd rails because they have significantly different sectional and material composition properties. Therefore rail weight was used as another criterion in determining track homogeneity. Finally, transposed rails and nontransposed rails were also included in the criteria. Transposing is the maintenance activity that interchanges the two rail legs to maximize rail life; therefore transposed sections may already be partially worn. The resulting homogeneous subnetworks under consideration are summarized in Figure 1. As mentioned earlier, the study considered only tracks in the Class A principal lines that carry annual tonnages of 1 to 5 million gross tonnes and curves that are less than 800 m in radius.

### Selection of Condition Bands

The condition parameter being predicted in the Markov model (i.e., rail wear) needs to be divided into a number of bands of condition. It is crucial that rail wear bands be selected carefully to coincide with renewals and maintenance triggers (threshold levels). In this way, the model can assign the required treatment once a proportion of the network reaches particular trigger levels. The treatments of interest in this study are rail renewal and rail transposing. The New Zealand track standards specify rail wear limits with regard to rail renewals and transposing, as described below (32).

#### 50-kg/m Rail Wear Limits

- If top wear only, rails need to be replaced when top wear reaches 18 mm, given there is no side wear;
- If both top and side wear are present, the maximum top wear is 14 mm and maximum side wear is 16 mm; rail has reached its wear limit when either top or side wear reaches these limits; and
- Welded rails can be transposed before top wear reaches a maximum of 12 mm or side wear reaches a maximum of 16 mm.

#### 91- and 90-lb/yd Rail Wear Limits

- Rails must be replaced on reaching 20 points in total (sum of top wear, side wear, and field wear) and
- Rails must be transposed on reaching 14 points in total.

The above rail wear thresholds, as specified in the track standards, needed to be reflected in the selection of rail wear bands to predict renewals and transposing requirements with the use of the Markov model. Given that Markov models do not retain the location reference, a challenge arose when the model for 50-kg/m rails was developed, for which there is a combined criterion of top wear and side wear (the standard states “with top wear between 6 and 14 points, maximum allowable combined wear is 16 points”). In practice, it was found from the data that in 99% of the records, the top wear on the high leg was less than 5 mm. In addition, in 97% of records the side wear on the low leg was 0 mm. Therefore, it was concluded that the combined criterion of top wear and side wear in 50-kg/m rails hardly ever triggers track maintenance activities. This conclusion was verified by track management staff and the combined criterion of top wear and side wear was subsequently removed from the analysis.

Because of the different treatment trigger levels for side wear and top wear in 50-kg/m rails and 91- and 90-lb/yd rails, two different sets of bands were considered. All bands coincided with the various treatment triggers associated with the different rail types and wear parameters, as per the track standards. Table 1 contains the bands established for the 50-kg/m rails.

Renewal and transposing criteria for 91- and 90-lb/yd rails are based on the total wear count. Therefore a single set of bands was adequate for HLT and LLT. Table 2 contains the bands established for 91- and 90-lb/yd rails.

### Development of Initial Condition Vectors

The ICVs describe the initial condition distribution of the network. These vectors contain the proportions or percentages of the network in each band of condition at the start of the analysis period and are determined relatively simply from the current network condition. The proportions of the network are determined from the lengths of

TABLE 1 Rail Wear Bands for 50-kg/m Rails

Condition Band	HS (mm)	LT (mm)
1	0–4	0–3
2	5–8	4–7
3	9–12	8–11
4	13–15	12–17
5	≥16	≥18

**TABLE 2 Rail Wear Bands for 91- and 90-lb/yd Rails**

Condition Band	HLT (points)	LLT (points)
1	6–9	6–9
2	10–13	10–13
3	14–16	14–16
4	17–19	17–19
5	≥20	≥20

track associated with each rail wear band divided by the total length of the tracks in each data set.

As an example, the ICVs developed in preparation for the validation exercise are included in Tables 3 and 4, for 50-kg/m and 91- and 90-lb/yd rails, respectively. From the main data set, four data subsets were separated out, corresponding to each of the homogeneous subnetworks defined in Figure 1. Two ICVs were developed from each of those data subsets (for HS and LT in 50-kg/m rails and for HLT and LLT in 91- and 90-lb/yd rails), which generated eight corresponding ICVs.

### Development of Transition Probability Matrices

The TPM is the controlling parameter of a Markov chain and therefore needs to be determined as accurately as possible. Traditionally, there are two methods of determining the probabilities in TPMs, as follows:

- Using subjective opinion and
- Using historical data.

The first method of determining the probabilities in TPMs is by way of acquiring expert engineering opinion. This method is usually carried out when there are insufficient amounts of historical data available, which is all too common with infrastructure assets. The second method, adopted in this research, involves determining probabilities from observed historical data. The probabilities of moving from condition  $i$  to condition  $j$  in 1 year ( $p_{ij}$ ) are determined from Equation 6.

$$p_{ij} = \frac{N_{ij}}{N_i} \quad (6)$$

$N_{ij}$  is the number of rail sections in the homogeneous subnetwork that moved from condition  $i$  to condition  $j$  in 1 year, and  $N_i$  is the total number of rail sections in that homogeneous subnetwork that started the year in condition  $i$ . The proportions are likely to vary from year to year; therefore, more than 1 year of data is necessary to increase the accuracy of the model.

The rail wear records from Class A principal lines taken during the analysis period resulted in approximately 3,700 records for 50-kg/m rails and approximately 1,500 records for 91- and 90-lb/yd rails. These data sets were further subdivided according to whether the rails had been transposed or not transposed, and resulted in the following data sets:

- 50-kg/m nontransposed rails,
- 50-kg/m transposed rails,
- 91- and 90-lb/yd nontransposed rails, and
- 91- and 90-lb/yd transposed rails.

The  $N_i$  and  $N_{ij}$  for Equation 6 were determined from the data sets by using a macro program that accounted for the spatial and temporal

**TABLE 3 50-kg/m Rail ICVs Developed for Model Validation**

Condition Band	High-Leg Side Wear				Low-Leg Top Wear			
	Nontransposed		Transposed		Nontransposed		Transposed	
	Length (km)	Percent	Length (km)	Percent	Length (km)	Percent	Length (km)	Percent
1	365.270	83.61	15.017	53.25	414.485	94.88	20.853	73.95
2	48.807	11.17	6.899	24.47	19.148	4.38	6.503	23.06
3	18.607	4.26	4.932	17.49	2.953	0.68	0.843	2.99
4	4.173	0.96	1.351	4.79	0.271	0.06	0.000	0.00
5	0.000	0.00	0.000	0.00	0.000	0.00	0.000	0.00

**TABLE 4 91- and 90-lb/yd Rail ICVs Developed for Model Validation**

Condition Band	High-Leg Total Wear				Low-Leg Total Wear			
	Nontransposed		Transposed		Nontransposed		Transposed	
	Length (km)	Percent	Length (km)	Percent	Length (km)	Percent	Length (km)	Percent
1	178.423	63.97	8.873	50.04	245.155	87.90	3.769	21.26
2	76.859	27.56	5.662	31.93	26.665	9.56	5.038	28.41
3	18.619	6.68	2.296	12.95	4.405	1.58	6.899	38.91
4	4.221	1.51	0.254	1.43	2.119	0.76	0.856	4.83
5	0.794	0.28	0.647	3.65	0.572	0.21	1.170	6.60

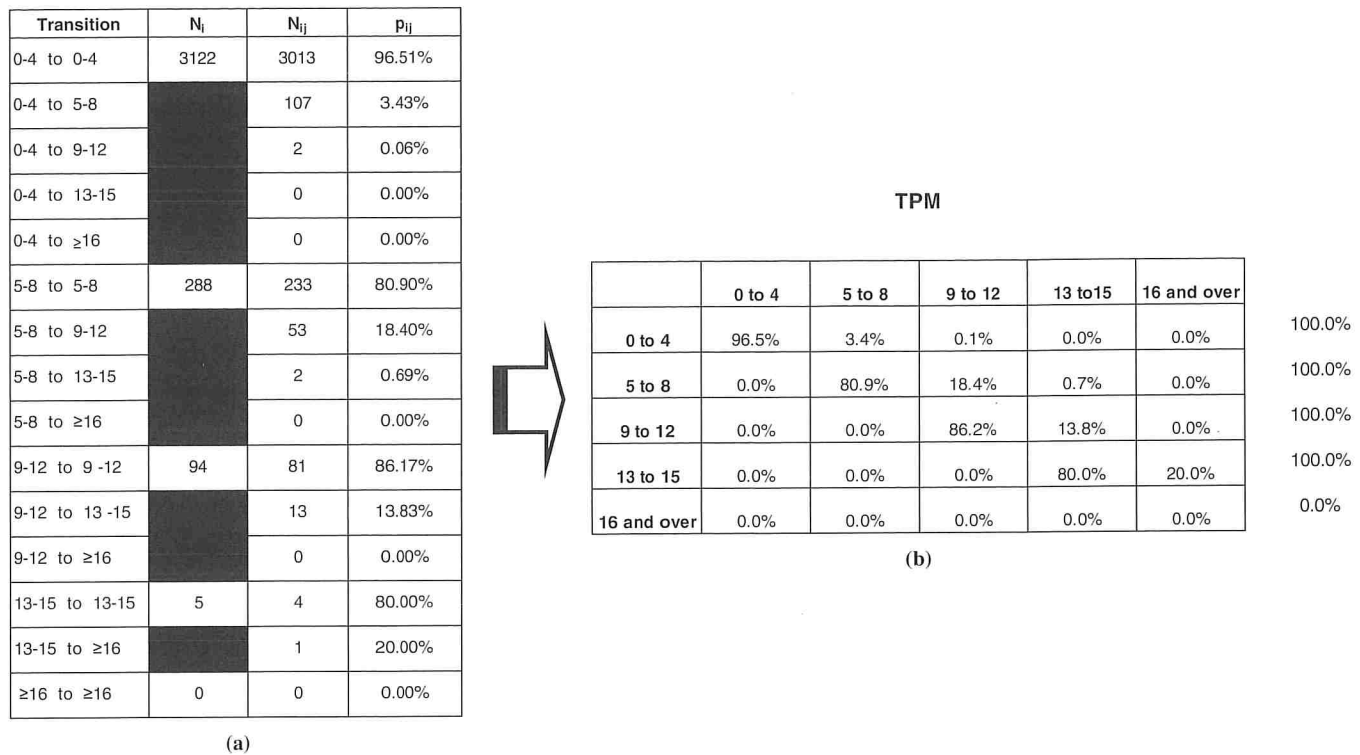


FIGURE 2 Calculation of TPM probabilities: (a) tabulation of counts for  $N_i$  and  $N_{ij}$  and calculated transition probabilities and (b) standard TPM format.

nature of the data. The macro also discounted data from locations where rail replacement or transposing occurred during the intervening period. At the end of analyzing all the records, the counts for  $N_i$  and  $N_{ij}$  were tabulated as shown in the example in Figure 2a. The transition probabilities were then calculated by using Equation 6 and are recorded in the  $p_{ij}$  column of Figure 2a. These probabilities were then rearranged into the standard format of a TPM as shown in Figure 2b. This same procedure was carried out for each homogeneous subnetwork, resulting in eight TPMs corresponding to the eight ICVs in Tables 3 and 4.

The TPMs developed with Equation 6 resulted in some probabilities not being determined because of insufficient data to calculate the particular probabilities. Incomplete TPMs are expected especially when probabilities for cells near the bottom of each TPM are calculated. The reason is that, generally speaking, there are fewer rail sections starting in higher rail wear bands (cells near the bottom of the matrix). As a further complication, in a number of cases the rail sections that did start in those bands did not have a successive reading taken, as they were replaced in the intervening period. Consequently, some of the transition probabilities have readings of 0%, indicating simply that there were no data points from which to determine those probabilities; see, for example, the 0% probability in the rightmost cell in the bottom row in Figure 2b.

For the model to function properly, the probabilities in each row of the TPM must add up to 100%, and at least two cells in each row must have non-zero probabilities. This is, of course, with the exception of the rightmost cell in the bottom row, which must always be 100%. If there was only one probability (i.e., 100% in one cell in a row), the model would not be able to simulate the deterioration of the network beyond that band. The TPMs there-

fore needed revision, with missing probabilities being estimated by using engineering judgment.

This process was carried out by assessing each TPM individually and examining the probabilities calculated from the rail wear data. A good indication of appropriate probabilities for the missing cells can be determined by comparing them with other calculated probabilities in the TPM. When all TPMs were filled with appropriate probabilities, a final check was also done by comparing all TPMs together, to ensure that they all contained sensible probabilities. Finally, all probabilities in the TPMs were rounded to whole numbers. Because this is a network-level simulation model, there is little advantage in retaining decimal points in the probabilities. The final TPMs to be used in Markov chains are presented in Tables 5 and 6, for 50-kg/m and 91- and 90-lb/yd rails, respectively.

## DISCUSSION AND CONCLUSION

Essentially all models are wrong but some models are useful (33). The means of identifying those models that are useful is by way of a suitable form of validation. In deterioration models, this validation is generally carried out by comparing the actual performance during a period of time, against the model-predicted performance for the same period of time.

The track database contained 10 years of historical data from which to develop the model, with the TPMs determined from the first 5 years of data. By using the ICVs and TPMs developed in this paper and substituted into Equation 4, along with information from the track database on renewals and transposing activities, the model can be validated with data from the remaining 5 years. The validation

TABLE 5 50-kg/m Rail TPMs

Modeled Wear Type	Starting Condition Band	Finishing Condition Band (%)				
		1	2	3	4	5
High-leg side wear (nontransposed)	1	95	4	1	0	0
	2	0	80	19	1	0
	3	0	0	86	14	0
	4	0	0	0	80	20
	5	0	0	0	0	100
High-leg side wear (transposed)	1	84	15	1	0	0
	2	0	67	30	3	0
	3	0	0	65	35	0
	4	0	0	0	60	40
	5	0	0	0	0	100
Low-leg top wear (nontransposed)	1	99	1	0	0	0
	2	0	95	5	0	0
	3	0	0	95	5	0
	4	0	0	0	99	1
	5	0	0	0	0	100
Low-leg top wear (transposed)	1	92	8	0	0	0
	2	0	97	3	0	0
	3	0	0	98	2	0
	4	0	0	0	98	2
	5	0	0	0	0	100

TABLE 6 91- and 90-lb/yd Rail TPMs

Modeled Wear Type	Starting Condition Band	Finishing Condition Band (%)				
		1	2	3	4	5
High-leg total wear (nontransposed)	1	90	9	1	0	0
	2	0	90	9	1	0
	3	0	0	87	10	3
	4	0	0	0	90	10
	5	0	0	0	0	100
High-leg total wear (transposed)	1	83	17	0	0	0
	2	0	89	11	0	0
	3	0	0	85	15	0
	4	0	0	0	85	15
	5	0	0	0	0	100
Low-leg total wear (nontransposed)	1	96	3	1	0	0
	2	0	93	5	2	0
	3	0	0	81	16	3
	4	0	0	0	85	15
	5	0	0	0	0	100
Low-leg total wear (transposed)	1	66	21	11	2	0
	2	0	82	18	0	0
	3	0	0	93	7	0
	4	0	0	0	90	10
	5	0	0	0	0	100

exercise for the rail wear TPMs is currently being written and will be reported in subsequent publications.

## FURTHER RESEARCH

Further research is planned to develop TPMs by using the Transition Matrix Calculator developed by Costello et al. (34) based on research by Ortiz-Garcia et al. (35). This development may overcome some of the issues with missing probabilities in the higher rail wear bands.

## ACKNOWLEDGMENTS

The authors acknowledge the funding provided by the Foundation for Research Science and Technology, New Zealand Railways Corporation, and Transfield Services.

## REFERENCES

1. Premathilaka, A. S., S. B. Costello, and R. C. M. Dunn. Development of a Deterministic Rail Wear Prediction Model. *Road and Transport Research*, Vol. 19, No. 1, 2010, pp. 40–50.
2. Costello, S. B., M. S. Snaith, H. G. R. Kerali, V. T. Tachtsi, and J. J. Ortiz-Garcia. Stochastic Model for Strategic Assessment of Road Maintenance. *Proceedings of the Institution of Civil Engineers, Transport*, Vol. 158, No. 4, 2005, pp. 203–211.
3. Zobory, I. Prediction of Wheel/Rail Profile Wear. In *Vehicle System Dynamics—Proc., of the 1997 International Association for Vehicle System Dynamics*, IAVSD, Vol. 28, No. 2/3, Aug. 25–29, 1997, pp. 221–259.
4. Szabo, A., and I. Zobory. On Deterministic and Stochastic Simulation of Wheel and Rail Profile Wear Process. *5th Mini Conference on Vehicle System Dynamics, Identification and Anomalies*, Technical University of Budapest, Hungary, 1996.
5. Zobory, I., and A. Szabo. Probability Theory Based Analysis of Wheel and Rail Wear Phenomena on a Railway Network. *8th Mini Conference on Vehicle System Dynamics, Identification and Anomalies*, Technical University of Budapest, Hungary, 2002.
6. Thelen, G. A., and P. M. Lovette. A Parametric Study of the Lubrication Transport Mechanism at the Rail-Wheel Interface. *Wear*, Vol. 191, No. 1/2, 1996, pp. 113–120.
7. Sims, R. D., K. A. Miller, and G. F. J. Schepmann. Rail Lubrication Measurement. *Proc., 1996 ASME/IEEE Joint Railroad Conference*, Oakbrook, Ill., April 30–May 2, 1996.
8. Mutton, P. J., and C. J. Epp. Factors Influencing Rail and Wheel Wear. *Railway Engineering Symposium: Upgrading of Australia's Rail Transport Systems*. Preprints of papers. Melbourne, Australia, 1983.
9. *Railnet Code—Code of Special Instructions: Engineering Services*. Trans Rail Limited, Wellington, New Zealand, 2002.
10. Detwiler, P. O., and M. L. Nagurka. *Track Geometry Modeling for Rail Vehicle Studies*. American Society of Mechanical Engineers, Dynamic Systems and Control Division (Publication) DSC, 1985.
11. Brough, M., A. Stirling, G. Ghataora, and K. Madelin. Evaluation of Railway Trackbed and Formation: A Case Study. *NDT & E International*, Vol. 36, No. 3 SPEC, 2003, pp. 145–156.
12. Oscarsson, J. Dynamic Train-Track Interaction: Variability Attributable to Scatter in the Track Properties. *Vehicle System Dynamics*, Vol. 37, No. 1, 2002, pp. 59–79.
13. Shahin, M. Y. *Pavement Management for Airports, Roads, and Parking Lots*. Chapman and Hall, New York, 1994.
14. Howard, R. A. *Dynamic Probabilistic Systems*. John Wiley & Sons, New York, 1971.
15. Black, M., A. T. Brint, and J. R. Brailsford. Comparing Probabilistic Methods for the Asset Management of Distributed Items. *Journal of Infrastructure Systems*, Vol. 11, No. 2, 2005, pp. 102–109.
16. Jackson, N. C., R. Deighton, and D. L. Huft. Development of Pavement Performance Curves for Individual Distress Indexes in South Dakota Based on Expert Opinion. In *Transportation Research Record 1524*, TRB, National Research Council, Washington, D.C., 1996, pp. 130–136.
17. Golabi, K., and R. B. Kulkarni. Arizona's Statewide Pavement Management System. *Civil Engineering*, Vol. 53, No. 3, 1983, pp. 43–47.
18. Kulkarni, R. B. Dynamic Decision Model for a Pavement Management System. In *Transportation Research Record 997*, TRB, National Research Council, Washington, D.C., 1984, pp. 11–18.
19. Butt, A. A., M. Y. Shahin, K. J. Feighan, and S. H. Carpenter. Pavement Performance Prediction Model Using the Markov Process. In *Transportation Research Record 1123*, TRB, National Research Council, Washington, D.C., 1987, pp. 12–19.
20. Thompson, P. D., L. A. Neumann, M. Miettinen, and A. Talvitie. A Micro-Computer Markov Dynamic Programming System for Pavement Management in Finland. *Proc., 2nd North American Conference on Managing Pavements*, Toronto, Ontario, Canada, Nov. 2–6, 1987.



21. Kerali, H. R., and M. S. Snaith. *NETCOM—The TRL Visual Condition Model for Road Networks*. Digest of Contractor Report 321. Transport Research Laboratory and the University of Birmingham, United Kingdom, 1992.
22. Wang, K. C. P., J. Zaniewski, and W. George. Probabilistic Behavior of Pavements. *Journal of Transportation Engineering*, Vol. 120, No. 3, 1994, pp. 358–375.
23. Li, N., W.-C. Xie, and R. Haas. Reliability-Based Processing of Markov Chains for Modeling Pavement Network Deterioration. In *Transportation Research Record 1524*, TRB, National Research Council, Washington, D.C., 1996, pp. 203–213.
24. MacLeod, D. R., and R. Walsh. *Report on Markov Modelling—A Case Study. Ontario and Yukon*. Public Works and Government Services Canada and Yukon Department of Community and Transportation Services, Yukon, Canada, 1996.
25. Tack, J. N., and Y. J. Chou. Pavement Performance Analysis Applying Probabilistic Deterioration Methods. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 1769, TRB, National Research Council, Washington, D.C., 2001, pp. 20–27.
26. Chung, K. L. *Markov Chains with Stationary Transition Probabilities*. Springer, Berlin, New York, 1967.
27. Isaacson, D. L., and R. W. Madsen. *Markov Chains—Theory and Applications*. Wiley, New York, 1976.
28. Cook, W. D., and A. Kazakov. Pavement Performance Prediction and Risk Modelling in Rehabilitation Budget Planning: A Markovian Approach. *Proc., 2nd North American Conference on Managing Pavements*, Toronto, Ontario, Canada, Nov. 2–6, 1987.
29. Butt, A. A., M. Y. Shahin, S. H. Carpenter, and J. V. Carnahan. Application of Markov Process to Pavement Management Systems at Network Level. *Proc., 3rd International Conference on Managing Pavements*, San Antonio, Tex., 1994.
30. Lytton, R. L. Concepts of Pavement Performance Prediction Modeling. *Proc., 2nd North American Conference on Managing Pavements*, Toronto, Ontario, Canada, Nov. 2–6, 1987.
31. Micevski, T., G. Kuczera, and P. Coombes. Markov Model for Storm Water Pipe Deterioration. *Journal of Infrastructure Systems*, Vol. 8, No. 2, 2002, pp. 49–56.
32. *Infrastructure Engineering Handbook—T 200: Engineering Services*. Tranz Rail Limited, Wellington, New Zealand, 2000.
33. Box, G. E. P. Robustness in the Strategy of Scientific Model Building. In *Robustness in Statistics* (R. L. Launer and G. N. Wilkinson, eds.), Academic Press, New York, 1979.
34. Costello, S. B., J. J. Ortiz-Garcia, and M. S. Snaith. Analytical Tool for Calculating Transition Probabilities for Pavement Performance Prediction. Presented at 90th Annual Meeting of the Transportation Research Board, Washington, D.C., 2011.
35. Ortiz-Garcia, J. J., S. B. Costello, and M. S. Snaith. Derivation of Transition Probability Matrices for Pavement Modeling. *Journal of Transportation Engineering*, Vol. 132, No. 2, 2006, pp. 141–161.

---

*The Railway Maintenance Committee peer-reviewed this paper.*

# Asset Condition Assessment at Regional Transportation Authority in Chicago, Illinois

Grace Gallucci, John Goodworth, and John G. Allen

Traditionally, transit systems managed their assets on an as-needed basis, but the urgency of overcoming deferred maintenance in the 1980s brought about an interest in capital programming to systematize the modernization of the physical plant better. Since the early 21st century, efforts to plan for capital maintenance needs have taken the form of the more ongoing process of asset management. Responding to several critical needs in the transit infrastructure of the Chicago, Illinois, area since the late 1980s, the Regional Transportation Authority (RTA) commissioned an asset condition assessment for transit in northeastern Illinois. The resulting study, released in 2010, is an industry leader in taking stock of an entire region's transit capital assets. The study establishes an estimated replacement cost for the system and 10-year state-of-good-repair needs. RTA intends to make an annual revision of this assessment into an ongoing process and is using the data as part of its overall performance measurement process. A decision tool is being developed in cooperation with other transit systems to help prioritize investments on the basis of the condition, importance, and impact of different assets.

The early 21st century finds transit agencies struggling to accomplish more with fewer resources. As a new generation of transportation officials faces up to the need to rehabilitate or replace the aging rolling stock and deteriorating physical plants, it becomes all the more important to invest limited capital funds where they will do the most good. To best prioritize needs, the Regional Transportation Authority (RTA) of northeastern Illinois is systematically taking stock of the physical assets of the transit operators serving the Chicago, Illinois, metropolitan area.

This paper first considers the historical background of today's asset condition assessment efforts, which are a modern outgrowth of efforts to reverse the long-standing physical neglect of major transit systems—an unfortunate, painful, and expensive process that came to a head in the 1970s and led to extensive rehabilitation work starting in the 1980s.

Next, the Chicago institutional setting is briefly described, followed by an exploration of situations that have arisen since the 1970s in which deteriorated rail lines were rebuilt. The RTA asset condition assessment study is then described, followed by a discussion of the involvement of the FTA and future directions for the study.

G. Gallucci, Suite 1650, and J. Goodworth and J. G. Allen, Suite 1550, Regional Transportation Authority, 175 West Jackson Boulevard, Chicago, IL 60604-2705. Corresponding author: J. G. Allen, allenj@rtachicago.org.

*Transportation Research Record: Journal of the Transportation Research Board*, No. 2289, Transportation Research Board of the National Academies, Washington, D.C., 2012, pp. 111–118.  
DOI: 10.3141/2289-15

## CHALLENGE OF DETERIORATING ASSETS

Through the 1970s, transit authorities had largely been upgrading their physical plants in response to pressing needs. Equipment or facilities at the end of their useful service lives were replaced if the agency had the requisite combination of resources and political support. Otherwise, the assets were allowed to decay, sometimes almost to the point of becoming unusable.

By the late 1970s the limitations of this ad hoc approach to the long-run upkeep of buses, rail cars, and fixed facilities became painfully evident on systems such as Ohio's Greater Cleveland Regional Transit Authority (1), Philadelphia's Southeastern Pennsylvania Transportation Authority (SEPTA) (2–4), and most notably the New York City Transit Authority (5, 6). Although maintenance standards varied among properties, by the end of the 1970s practically every system with tracks, structures, and stations dating from the years between the 1890s and the 1930s was experiencing capital maintenance issues. This description of Philadelphia's commuter rail system could apply equally to several other rail systems in older industrial cities, where “much of the physical plant was permitted to decay. Ridership was dropping fast after the boom years of World War II, and expenses were rising even faster. . . . Routine maintenance of stations all but ceased. Even such things as broken steps and burned-out light bulbs were often ignored. . . . Decreased service and frequent delays and breakdowns resulted in more and more riders seeking alternative means of getting to work, usually turning to their automobiles” (7, p. 107).

Eventually matters came to a head as transit authorities faced critical decisions about rehabilitating key rail lines. New York City Transit Authority (now New York City Transit), SEPTA, Chicago Transit Authority (CTA), and other properties found themselves spending huge sums to thoroughly rebuild and rehabilitate tracks, structures, stations, and other assets, in large measure because they had been allowed to fall well below a state of good repair. On Chicago's Green Line (rebuilt from 1994 to 1996), “A century of frequent and continuous service, much of it by trains composed of heavy-weight cars, did not permit the steel spans to age gracefully. . . . In addition, the track had far exceeded its useful life. . . . The combined impact of these deficiencies caused drastic reductions in speed limits across large segments of the line. Running time suffered accordingly. For example, the running time . . . between Oak Park and the [downtown] Loop, was 20 minutes in 1983. By 1994, the running time had increased to over 30 minutes” (8, pp. 108, 109).

Fortunately, most disruptions resulting from deferred maintenance have been confined to scheduled downtime for reconstruction work, but occasionally events have intervened unexpectedly. In 1977 cracks on a rapid transit bridge temporarily forced suspension

of service on the CTA's busy Dan Ryan rapid transit line and the Rock Island commuter rail line running beneath the bridge (9). In another instance, a week after the 1984 opening of Philadelphia's long-awaited Center City commuter rail tunnel, SEPTA was forced to suspend service on half of the combined operation when structural flaws were found on one part of the former Reading Company's viaduct through North Philadelphia. Repairs were made and service was restored after 17 days, but the issues were not limited to that particular segment of the viaduct. SEPTA conducted a detailed engineering study and completely rebuilt the viaduct, which required prolonged track outages during the summers of 1992 and 1993 (10).

The accumulation of deferred maintenance does not affect the ability of transit riders to complete trips successfully. Deferred maintenance can, however, significantly affect travel times as transit agencies impose slow zones along poorly maintained sections of track to maintain safe operations. Furthermore, it can also affect the reliability of travel, as unexpected delays (particularly those due to mechanical failures of vehicles or infrastructure systems) force riders to add greater time margins when planning travel than would be the case for a transit property performing more comprehensive maintenance on its rolling stock and fixed plant.

### CAPITAL PROGRAMMING: ASSET MANAGEMENT PRECURSOR

By the early 1980s the earlier approach of reinvesting in the system only when the need became conspicuously apparent had become discredited. The sheer magnitude of investment needed in New York's far-flung but neglected system showed that individual system elements could no longer be considered in isolation from one another.

As transit agencies sought to address their sometimes daunting maintenance and replacement needs, they developed a more systematic approach, which became known as capital programming. A major part of capital programming involved using information technology to inventory a property's assets and monitor the progress of rehabilitation and replacement projects. The New York City Transit Authority and its parent agency, the Metropolitan Transportation Authority (MTA), were leaders in this process.

Capital programming combined elements of strategic planning and project management, but the most important advance it offered was to apply a systems approach to all the capital assets of a transit system. One capital programming effort involving seven medium-to larger-size properties was described in these terms: "[T]he San Francisco [California] area Metropolitan Transportation Commission [MTC] . . . initiated a project . . . to set regional priorities for capital investment. Borrowing a concept from private sector strategic planning, the MTC undertook . . . to provide the region with a preliminary estimate of its *capital readiness* to maintain and enhance the public transportation system . . . over the long term" (11, p. 1).

In 1988 TRB devoted *Transportation Research Record 1165* entirely to capital programming and strategic planning. In that volume, papers discussed applications of capital programming at Chicago's RTA, the Port Authority Trans-Hudson in New Jersey, the San Diego, California, Metropolitan Transportation Development Board (the functions of which have since been reallocated to other agencies), the King County Metro in Seattle, Washington, and the Washington Metropolitan Area Transit Authority in Washington, D.C.

## RISE OF TRANSIT ASSET MANAGEMENT

With transit agencies putting their physical assets in a significantly better state of repair by the early 1990s, capital programming seems to have receded somewhat from the industry's consciousness. But a new set of urgent state-of-good-repair needs for rehabilitating aging infrastructure has since emerged, including Philadelphia's Frankford and Market Street rapid transit lines and several Chicago rail lines (12, 13).

A new generation of transit officials became interested in asset management, as industry leaders became accustomed to the fact that keeping a large and complex transit property in a state of good repair requires ongoing vigilance and work. Up through the late 1990s, "asset management was something private-sector companies did. . . . In September 1996, AASHTO and FHWA held the first asset management workshop focused on sharing experiences in the public and private sectors. . . . The first [AASHTO-FHWA] workshop on asset management defined [it] as 'a systematic process of maintaining, upgrading, and operating physical assets cost-effectively. It combines sound business practices and economic theory, and it provides tools to facilitate a more organized logical approach to decision making'" (14, p. 21).

Asset management is closely related to capital programming of the 1980s in its concern with catching up with and keeping ahead of maintenance needs, but it is more immediately concerned with monitoring the physical condition of buses, trains, stations, platforms, tracks, rights-of-way, garages, shops, yards, and transit terminals. There is also a greater interest in private-sector criteria for assessment and decision making. Since the 2000 publication of *Transportation Research Record 1729*, which included several articles on transportation asset management, a substantial body of literature has emerged on asset management for transportation (15–22).

Transit systems have been part of this new emphasis on asset management (23). New York's MTA and the San Francisco MTC have assessed their long-term capital needs (24, 25). During the early years of the 21st century, the CTA was in the forefront of using information technology to document the condition of its rail and bus assets (26, 27).

### TRANSIT ASSET MANAGEMENT IN CHICAGO

RTA is the region's lead agency for asset management at one of the largest transit networks in the United States. RTA is the funding, oversight, and long-range planning agency for the three following transit operators, known in northeastern Illinois as service boards:

- CTA operates bus and rapid transit in the city of Chicago and several suburbs in adjacent parts of Cook County (of which Chicago is the county seat);
- Metra operates commuter rail throughout the six-county northeastern Illinois region; and
- Pace operates fixed-route bus service throughout the suburban parts of the region (also serving rapid transit stations and other connection points with CTA in and adjacent to the city of Chicago). Pace is also the paratransit agency for all parts of the region, city and suburbs.

### BACKGROUND FOR ASSET MANAGEMENT IN CHICAGO TRANSIT

There is ample justification for assessing the condition of the transit system in northeastern Illinois. Transit agencies in the Chicago area have reinvested heavily in the physical future of rail lines and have

made choices among alternative construction plans with varying effects on cost levels, the duration of construction, service provided during the construction period, and (indirectly) on ridership.

- By the late 1970s the commuter rail operation of the bankrupt Chicago, Rock Island and Pacific Railroad was in a very poor state of repair, with substantial effects on running times and reliability. Even though slow orders were widespread, low-speed derailments were common enough to be unremarkable. Starting in 1978 (before Metra's formation as a separate agency in 1983) RTA oversaw a complete physical overhaul of the Rock Island's commuter territory, now owned, operated, and maintained by Metra (28, 29).

- CTA closed the Skokie Swift (today's Yellow Line) between Howard and Skokie–Dempster for thorough rebuilding of the track and roadbed between July and November 1991. Although the Yellow Line was built to very high standards when it opened in 1925, normal wear and tear during the intervening decades (resulting in slow orders that affected operations) made the complete physical renewal of the line necessary. CTA chose an extended closure with nonstop substitute bus service as less problematic than attempting to operate trains reliably with single-tracking, particularly during peak hours (8).

- By the early 1990s, CTA's Green Line, comprising the Lake Street "L" (as elevated lines are known in Chicago) on the West Side and the South Side "L," had deteriorated to the point at which the Green Line, which was greatly affected by slow orders, was barely suitable for operation. CTA was weighing investment options on all three of its West Side rail lines when a decision was made to rebuild the original elevated structures (30). The Green Line was closed altogether for reconstruction in January 1994, reopening in May 1996 (31, 32). CTA provided substitute express bus service during the closure (33). By giving contractors exclusive occupancy for slightly more than 2 years, the total duration of construction was minimized. However, some riders who made satisfactory alternative arrangements while service was suspended did not return to the Green Line when it reopened.

- A decade later, the Douglas Park "L," then known as the 54–Cermak branch of the Blue Line and today operated as CTA's Pink Line, had deteriorated to the point at which it was necessary to replace the original elevated structures and stations with aerial guideways and new stations. To avoid prolonged closures, construction crews took possession of the line only on weekends. Although this weekend work resulted in a longer and more expensive construction period (from 2001 through 2005), CTA did not face the same challenge of rebuilding ridership as it did with the reopened Yellow and Green Lines.

- CTA performed extensive reconstruction work on the Dan Ryan portion of the Red Line—the busiest rail line on the South Side—between 2004 and 2007. Opened in 1969, the line had sustained much wear and tear on its tracks and roadbed after three and a half decades of operation in the median strip of a busy expressway. To ensure continuous service, much temporary trackage had to be built (which resulted in operation at reduced speeds).

- In 1991 and again in 2007–2008, CTA performed extensive track work in the State Street Subway portion of the Red Line. The tight confines of the subway prevented this work from being done while trains were operating. CTA provided exclusive occupancy windows for construction on weeknights and weekends by routing Red Line trains onto the elevated alignment used by the Brown Line.

- Between 2007 and 2009, CTA carried out its Brown Line capacity enhancement project, which involved rebuilding stations so that they could accommodate trains of up to eight cars (rather than six)

and complied with the Americans with Disabilities Act. During the construction period, CTA needed to close one of the four tracks along the busy segment between Fullerton and Belmont on Chicago's North Side, which accommodates the Red Line and the peak period Purple Line Express as well as the Brown Line (34). To meet the needs of riders whose travel was disrupted, CTA added buses on several routes in its north lakefront corridor, and Metra added trains on its Union Pacific North Line.

- Deteriorating ties on the O'Hare extension of the Blue Line, opened in 1983 and 1984, resulted in extensive slow orders by the middle of the first decade of the 21st century (35). In 2008 CTA rebuilt the affected section, with extensive evening and weekend track closures and bus substitutions.

- Metra faces the need to rebuild aging, obsolete bridges carrying the Union Pacific North Line over 22 streets on Chicago's North Side. Metra's Union Pacific (UP) lines use the oldest locomotives in Metra's fleet because the North Side bridges cannot accommodate more modern, heavier locomotives. Because commuter train consists are interlined between the UP North, Northwest, and West Lines to use equipment as efficiently as possible, all three UP lines must now use the oldest type of locomotive (and will benefit from the ability to use newer locomotives once bridge reconstruction is completed). Metra has planned work so as to preserve two-track operation throughout the line during the 8-year construction period. Although this will require more time and money to implement than single-tracking, it will minimize disruption on Metra's fourth-busiest line (and the one with the greatest reverse commuting).

- One of CTA's current major investment studies is a study of the future of the Red and Purple Lines on Chicago's North Side and in Evanston, Illinois. An aging filled-earth embankment built between 1914 and 1922 carries the four tracks of the Red and Purple Lines between Lawrence (where a steel elevated structure ends and the embankment begins), Howard (the north end of the Red Line), and Evanston. CTA is seeking to determine what engineering alternative best ensures the future of this vitally important part of the network (36).

Pressing state-of-good-repair needs remain at CTA and the other service boards, and these needs sometimes make themselves known unexpectedly. A 2007 rapid transit derailment attributed to an undetected track segment that was out of gage caused both CTA and RTA (the state-designated rail safety oversight agency for CTA) to pay increasing attention to track inspection and other safety matters (37, 38). The two agencies now work together on rail safety issues in an atmosphere of cooperative problem solving (39).

Aligning CTA's internal processes more closely with safety needs has been helpful, but deferred maintenance remains a very real issue. This deferral has brought increased attention to the condition of the system.

## ASSET CONDITION STUDY

With much of the conversation about regional transit in northeastern Illinois involving the system's capital investment backlog, it made sense for RTA to commission an assessment of the system's physical plant and rolling stock as a prelude to determining how best to prioritize limited capital for investment. In 2007 RTA updated its strategic plan for the first time in more than a decade (40). Assessing the condition of the transit system's capital assets was a logical outgrowth of the strategic plan, as it was clear that a prerequisite for a system in a state of good repair was a clear understanding of just what shape that



system was in. Although RTA and the service boards knew what the system's assets were, there was no precise inventory in existence, as all three service boards were making changes to their fixed plant and rolling stock, sometimes on an improvised basis.

With the encouragement of FTA, RTA began a regional asset condition assessment in January 2009 and issued a report in August 2010, becoming one of the first transit agencies in the United States to conduct a study of this nature. The *Regional Transportation Authority Capital Asset Condition Assessment* inventoried and assessed, as best as could be done without performing detailed engineering studies of large parts of the system, what the region's transit assets were and their condition (41). All three service boards participated in all aspects of the study.

## Study Process

The first step was to record all the system's assets. To make sense of the many different types of rolling stock and fixed plant items, it was necessary to group the various assets into five broad categories:

- Track and structures;
- Electrical and subway equipment;
- Signals, communications, and fare collection;
- Stations, garages, and facilities; and
- Rolling stock.

Table 1 shows these categories in more detail.

The second step was to determine the condition of these assets. It was not feasible to establish the condition of all of the assets in civil engineering terms within the time and budgetary constraints of the study. Therefore, the age of the assets was used as a proxy for their condition. New assets were judged to be the highest rated, at "5—in excellent condition." Other assets were rated in descending order according to their age, in regard to the useful life of the particular type of asset. These ratings were "4—good," "3—fair," "2—poor," and "1—beyond useful service life." Assets in the latter category were judged to be in need of replacement and are characterized as "backlogged," that is, awaiting investment in the system's capital stock.

To verify that the age interpolation process was consistent with the system's true condition, a sampling was made of approximately 1% of the assets, chosen so as to include the most relevant major asset categories for each service board. The sampling process raised no serious questions about the validity of using age as a proxy for condition although there are necessarily exceptions on such large and complex systems as Chicago's. Some well-maintained older assets (such as Metra's rebuilt stainless-steel cars) may be in better condition than their age would suggest. Conversely, certain newer assets may have deteriorated faster than their years would indicate. The report, therefore, should not be taken as a precise assessment of the condition of specific assets. Instead, the study is most useful as an overview of the condition of the northeastern Illinois transit system in the aggregate.

Third, the replacement costs for the various assets were estimated. Input information for this process included purchase price, age, and depreciation rates. Professional judgment based on experience was used, along with research into the replacement costs of certain specific assets.

Fourth, replacement costs were determined for all assets that had reached life expiry. Fifth, 10-year normal replacement costs between 2010 and 2019 were established. Sixth, capital maintenance costs for the same period were established. Finally, these costs were added to produce a total 10-year state-of-good-repair need.

## Findings

By using a rating scale from 5 (excellent condition) to 1 (beyond useful service life), the state of the region's main groups of transit assets was found to be as follows:

- Rail passenger cars. A rating of 2.29 was achieved, with nearly 42% beyond their useful life. To bring the rail fleet into a state of good repair, 931 of the 2,225 cars would have to be replaced, and the remaining fleet would have to be maintained at a state of good repair. Figure 1 shows the categories for rail passenger cars.
- Rail stations. The rating of 3.00 belies the finding that more than 39% of the stations were rated at 1 (beyond useful service life). At least 150 of the region's 382 rail stations would have to be renovated, with appropriate replacement and capital maintenance performed at the other stations.
- Rail bridges and structures. The 3.26 rating appears fairly positive at first glance, but this figure conceals the fact that 11% of the bridges and structures are beyond their useful service life. At least 151 of the 1,361 bridges and structures would have to be renovated between 2010 and 2019 to bring them into a state of good repair, and the many remaining elements would have to be well maintained. Figure 2 shows the categories for rail bridges and structures. As an example of the differences in condition among assets of the same type, Figure 3 shows a rapid transit viaduct in excellent condition (rating of 5) with another beyond its useful service life (rating of 1).
- Fixed-route buses. The bus fleet overall was rated 3.46, with two-thirds in a state of good repair. This rating implies the need to replace at least 457 of the region's 2,918 buses and maintain the others at a state of good repair.
- Rail maintenance facilities. Rail car and locomotive shops were rated at 3.64 overall. Of these, 14% are rated at 1 (beyond their useful life) and another 8% are rated at 2 (marginal). To bring the rail maintenance facilities into a state of good repair, at least five of the 36 will have to be renovated and the others properly maintained.
- Bus maintenance facilities. An overall rating of 3.37 was achieved, with 16% rated beyond their service life and another 16% rated marginal. At least three of the 19 facilities will need to be replaced, in addition to maintaining the other garages at a state of good repair.

Several cost components were identified in conjunction with the system's capital stock:

- Backlog of assets that have passed their useful service lives and for which replacement is therefore indicated amounts to \$13.8 billion for the region (\$10 billion for CTA, \$3.7 billion for Metra, and \$100 million for Pace).
- Normal replacement costs for assets expected to reach the end of their service lives between 2010 and 2019 amount to \$6.8 billion regionwide (\$3.2 billion for CTA, \$1.7 billion for Metra, and \$1.9 billion for Pace).
- Capital maintenance component, that is, the cost of keeping assets in a state of good repair, comes to \$3.9 billion regionwide (\$1.7 billion for CTA, \$1.9 billion for Metra, and \$200 million for Pace).

Soft costs covering engineering, planning, and project management were added to the totals, as were contingency costs for unforeseen eventualities such as the discovery once work is under way that the original specifications for rehabilitation or replacement do not



TABLE 1 Types of Capital Assets Inventoried

Asset Group	Subgrouping	Specific Asset (Associated Transit Operators)
Track and structures	Track structures	Track structures (CTA, Metra)
	Track	Ties (CTA, Metra) Rail (CTA, Metra) Grade crossings (CTA, Metra) Special trackwork (CTA, Metra)
Electrical and subway equipment	Traction power	Substations (CTA, Metra) Substation distribution (CTA) Right-of-way traction power (CTA, Metra) Overhead catenary wire (Metra)
	Subway equipment	Subway electrical service (CTA) Subway fans (CTA) Subway illumination (CTA) Subway pumps (CTA)
Systems	Signals	Interlockings (CTA, Metra) Interlockings owned by railroads operating under purchase-of-service contracts (Metra) Cab signals (CTA) Signal controls (Metra) Signal controls owned by railroads operating under purchase-of-service contracts (Metra) Grade crossing signals (CTA, Metra) Grade crossing signals owned by railroads operating under purchase-of-service contracts (Metra)
	Fare collection communications	Fare collection equipment (CTA, Metra, Pace) Radio systems (CTA, Metra, Pace) GPS on-board bus (CTA) CCTV system at rail stations (CTA) Cable plant (CTA, Metra) Fiber optic backbone network (CTA, Metra) Station SCADA systems (CTA) Electrical substation SCADA RTUs (CTA) Public address systems—audio (CTA) Public address systems—vocal management system (CTA) CCTV system for ticket vending machines (Metra) CCTV system for homeland security (Metra) Telephone systems (Metra) Public address systems (Metra) Microwave (Metra) Wireless telephone (Metra) Electric signal—ITS (Pace)
Stations, garages, facilities	Stations and parking	Stations (CTA, Metra) Station parking (CTA, Metra)
	Passenger and maintenance facilities	Bus passenger facilities and stations (CTA, Pace) Bus garages (CTA) Other bus maintenance facilities (CTA) Bus support facilities and equipment (Pace) Paratransit support facilities and equipment (Pace) Rail maintenance facilities (CTA) Rail yards (CTA) Maintenance and yard facilities (Metra) Agency headquarters (CTA, Metra, Pace)
Rolling stock	Revenue vehicles	Electric multiple unit cars (CTA, Metra) Diesel locomotives (Metra) Coaches for diesel push-pull service (Metra) Buses (CTA) Nonparatransit vehicles (Pace) ADA service vehicles other than vans (Pace) ADA paratransit vans (Pace)
	Nonrevenue vehicles work equipment	Nonrevenue vehicles (CTA, Metra, Pace) Work equipment (CTA, Metra, Pace)

NOTE: ADA = Americans with Disabilities Act; CCTV = closed-circuit television; GPS = Global Positioning System; ITS = intelligent transportation systems; RTU = remote terminal unit; SCADA = supervisory control and data acquisition.

SOURCE: *Regional Transportation Authority Capital Asset Condition Assessment*, Aug. 2010, p. 7.

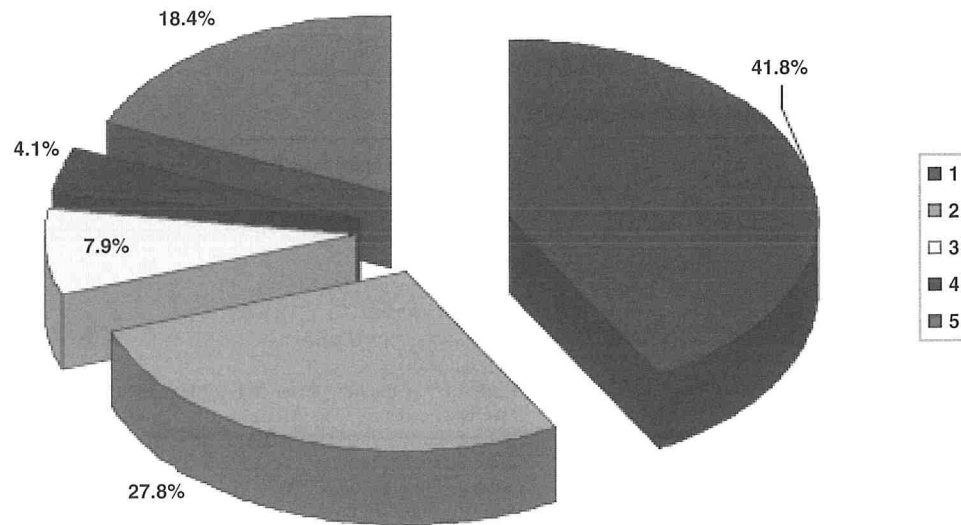


FIGURE 1 Condition of Metra and CTA rail passenger cars: 1-rating (beyond useful service life) starts at upper right, advancing clockwise to 5-rating (excellent condition) at upper left. (Source: *Regional Transportation Authority Capital Asset Condition Assessment*, Aug. 2010, p. vi.)

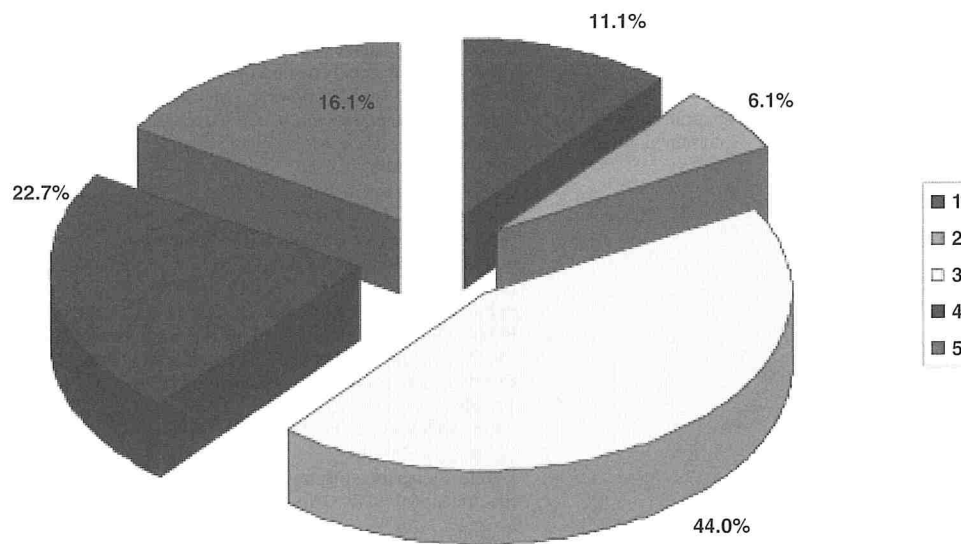


FIGURE 2 Condition of Metra and CTA bridges and structures: 1-rating (beyond useful service life) starts at upper right, advancing clockwise to 5-rating (excellent condition) at upper left. (Source: *Regional Transportation Authority Capital Asset Condition Assessment*, Aug. 2010, p. vi.)



(a)



(b)

FIGURE 3 Rapid transit viaducts: (a) in excellent condition (Church Street, Evanston, Ill., CTA Purple Line) and (b) beyond useful service life (Hollywood Avenue, Chicago, Ill., CTA Red Line) (Regional Transportation Authority photos).

meet project needs. All in all, the system's 10-year state-of-good-repair needs total \$24.6 billion, or \$2.46 billion annually. Of this amount considered necessary to bring the system up to a state of good repair and keep it there, CTA's needs amount to 61%, Metra's 30%, and Pace's 9%.

Even aside from the fact that assessing the condition of the physical plant is a de facto precondition for federal funding, RTA believes this course of action is both right and cost-effective. The study cost \$1.2 million to perform, and it will cost another \$300,000 to update the data during the next 5 years (just over 1/1000 of a percent of the annual cost of rehabilitating the system). The study gives RTA better information to prioritize projects within resource constraints. There are quantifiable benefits in the ability to reduce operating costs associated with deteriorating assets and in the ability to retain and attract customers when the system is in good condition.

These capital needs are reasonable compared with the net replacement value of the capital assets of CTA, Metra, and Pace throughout northeastern Illinois. Before the study, RTA had been estimating the system's replacement value at \$35 billion, but this now appears to fall far short of the true cost of rebuilding or replacing the entire physical plant. The study determined the net replacement value of the total capital assets used by the three service boards to be about \$140 billion in 2010 dollars. This replacement value includes an estimated \$100 billion for the structural components of the subways used by CTA's Red and Blue Line trains.

### Cooperation with the Federal Transit Administration

The RTA's asset condition assessment is relevant beyond northeastern Illinois. In 2008 FTA issued a report, *Transit State of Good Repair: Beginning the Dialogue*, in which FTA estimated that about a fourth of the capital assets of all U.S. transit properties are near or beyond the end of their useful service lives (42). A 2009 *Rail Modernization Study* examined the assets of seven rail transit agencies and found only 30% of their assets (weighted by value) to be in good or excellent condition. Fully 35% were in marginal or poor condition, and the remaining 35% were in adequate condition (43). FTA has remained an active participant in the conversation about transit state-of-good-repair issues (44).

As RTA's *Capital Asset Condition Assessment* was being funded by FTA, RTA decided to use the same categories and terms as FTA. This step helps the RTA assessment to fit into FTA's analytical framework, allowing FTA officials to evaluate RTA's study readily and see how it fits in with FTA's state-of-good-repair efforts. In some instances, similar assets are broken out differently for different service boards, but all assets are categorized according to FTA's framework.

RTA sees its asset condition assessment as an important analytical method worth sharing among the professional community. RTA delivered a presentation about its asset condition assessment at a July 20–22, 2011, workshop in Atlanta, Georgia, about state-of-good-repair issues, and at FTA's request, distributed copies of the assessment to representatives of other transit systems.

### Ongoing Process

Although the *Capital Asset Condition Assessment* is a major step toward guiding the Chicago area's transit system toward a state of

good repair, RTA views this step as just a beginning. In cooperation with FTA, RTA will be updating the assessment during the next 5 years. Each year engineering studies will be performed on more assets, which will help RTA update the estimated replacement value of the system and make the figure more precise, with less reliance on professional judgment (however experienced this may be) and more on physical inspection of assets.

In particular, RTA anticipates an improvement in the precision of the data on Metra's physical plant. As of this writing, Metra is arranging to bring in contractors to conduct an asset inventory, the results of which will help inform RTA's ongoing work.

Looking beyond the 5-year framework of the current asset condition assessment work plan, RTA seeks to make the updating process an ongoing part of its oversight role. Assets will continue to age and wear out, which makes asset condition management a vitally important prerequisite for any transit system seeking to reach (and stay at) a state of good repair.

RTA is also incorporating capital asset condition assessment data into its regional transit performance measures (45). RTA monitors, as part of a broader, ongoing process, service maintenance and capital investment indicators, including state-of-good-repair issues. At this writing, RTA is developing a performance measure for asset condition, based on the percent of assets in a state of good repair (3 or above in the rating scale), and is quantifying the cost of bringing the substandard elements up to a state of good repair. Given the need for rolling stock and fixed plant alike to support reliable operations, RTA sees this as an important aspect of the performance measurement process.

### From Inventorying to Prioritizing

However important it is for transit agencies to know about the condition of their assets, it is even more important for them to use this information to prioritize the use of scarce resources. RTA, New York's MTA, and the Los Angeles County Metropolitan Transportation Authority in California are currently working with a contractor to develop a decision tool for use in prioritizing investment in the system. The decision tool will use the asset condition assessment as its database. It will allow agencies to evaluate alternative uses for capital funds according to various criteria, such as maintaining safety, enhancing security, and reducing operating costs.

### CONCLUSION

An asset condition assessment such as that of RTA is a vitally necessary step in addressing state-of-good-repair challenges. The RTA assessment shows how to categorize and cost assets and determine the system's 10-year repair needs.

The ongoing interest in asset management at transit authorities reflects the fact that large transit properties, particularly historically established systems, must be brought up to and kept in a state of good repair to the maximum extent possible. As assets wear out, they need to be either rebuilt or replaced, and a failure to do so before the situation becomes serious may affect the system's operating reliability. In an institutionally complex setting, Chicago's RTA has taken the lead in assessing the condition of transit assets in northeastern Illinois to promote a regional conversation about prioritizing scarce capital to best address the system's many pressing needs.

## REFERENCES

1. Landgraf, R. J. Cleveland's Light Rail System in the 1980s: The Ongoing Revolution. In *Transportation Research Record 1361*, TRB, National Research Council, Washington, D.C., 1992, p. 259.
2. Kizzia, T. Philadelphia: Help Is on the Way—And None Too Soon. *Railway Age*, July 14, 1980.
3. Fahrenwald, B. Winning a Catch-Up Game. *Railway Age*, June 1984.
4. DeGraw, R. As Fast as a Speeding Bullet: Rebuilding the Norristown High-Speed Line. In *Transportation Research Record 1361*, TRB, National Research Council, Washington, D.C., 1992, pp. 272–275.
5. Gunn, D. L. New York's Transit Rebound. *Progressive Railroading*, Aug. 1989.
6. Hom, K. J. Reinventing Transit: The Twenty-Year Overnight Success Story of NYC Transit. Presented at the Metropolitan Conference on Public Transportation Research, Chicago, Ill., June 11, 1999.
7. DeGraw, R. Regional Rail: The Philadelphia Story. In *Transportation Research Record 1433*, TRB, National Research Council, Washington, D.C., 1994.
8. Abrams, S. H. CTA's Recent Experience with Major Rail Rehabilitation Projects: Construction Efficiency Versus Ridership Retention. In *Transportation Research Record 1623*, TRB, National Research Council, Washington, D.C., 1998, pp. 105–111.
9. Young, D. Disaster. *Mass Transit*, April 1978.
10. Vantuono, W. C. Meeting the RailWorks Challenge. *Railway Age*, March 1993.
11. *Transit Capital Planning in the San Francisco Bay Area*. Urban Mass Transportation Administration, U.S. Department of Transportation, Dec. 1982.
12. Palmer, J. W. Track Time: Construction or the Customer. In *Transportation Research Record 1623*, TRB, National Research Council, Washington, D.C., 1998, pp. 99–104.
13. Wolinsky, J. Detailed Planning and Careful Co-ordination Pays Off. *Metro Report International*, Dec. 2010.
14. McNeil, S., M. L. Tischer, and A. J. DeBlasio. Asset Management: What Is the Fuss? In *Transportation Research Record: Journal of the Transportation Research Board*, No. 1729, TRB, National Research Council, Washington, D.C., 2000, pp. 21–25.
15. Hamilton, W. E. Transportation: Asset Management. *Fiscal Forum*, Michigan House Fiscal Agency, Lansing, Feb. 2001.
16. Kleiner, Y. Scheduling Inspection and Renewal of Large Infrastructure Assets. *Journal of Infrastructure Systems*, Dec. 2001.
17. Hutchins, K. Managing Your Money: Asset Management Solutions. *Mass Transit*, Dec. 2005/Jan. 2006.
18. *Transportation Research Circular E-C076: Asset Management in Planning and Operations*. Transportation Research Board of the National Academies, Washington, D.C., June 2005.
19. *Transportation Research Circular E-C093: 6th National Conference on Transportation Asset Management*. Transportation Research Board of the National Academies, Washington, D.C., March 2006.
20. *Transportation Research Circular E-C131: Transportation Asset Management: Strategic Workshop for Department of Transportation Executives*. Transportation Research Board of the National Academies, Washington, D.C., Nov. 2008.
21. *Applying Transportation Asset Management in Connecticut*. Connecticut Academy of Science and Engineering, Hartford, Dec. 15, 2008.
22. *Model Framework for Assessment of State, Performance, and Management of Canada's Core Public Infrastructure*. National Research Council of Canada and National Roundtable on Sustainable Infrastructure, Ottawa, Ontario, Canada, May 2009.
23. Tomeh, O., S. Brady, and D. Skorupski. National Bus and Facilities Condition Assessment. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 1760, TRB, National Research Council, 2001, pp. 56–67.
24. *Twenty Year Capital Needs Assessment: 2010–2029*. Draft report. Metropolitan Transportation Authority, New York, Aug. 2009.
25. Capital Asset Inventory & State of Good Repair. Presentation to Board of Directors, San Francisco Municipal Transportation Authority, San Francisco, Calif., Aug. 3, 2010.
26. Shiffer, M., A. Chakraborty, B. Donahue, G. Garfield, R. Srinivasan, and S. McNeil. Spatial Multimedia Representation of Chicago Transit Authority Rail Infrastructure. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 1838, Transportation Research Board of the National Academies, Washington, D.C., 2003, pp. 1–10.
27. Zhai, D., M. Hämmerle, L. Râmasubramanian, D. Lise, J. Schniedel, and S. Abou-Sabbh. Managing the Chicago Transit Authority's Infrastructure Using Spatially Referenced Asset Management. In *Applications of Advanced Technologies in Transportation Engineering*, ASCE, Reston, Va., 2004.
28. Young, D. The Rock Island: Back From the Brink. *Mass Transit*, Nov. 1978.
29. Fahrenwald, B. Chicago: A Comeback Story? *Railway Age*, Feb. 1983.
30. Eash, R., K. Dallmeyer, and R. Cook. Ridership Forecasting for Chicago Transit Authority's West Corridor Project. In *Transportation Research Record 1402*, TRB, National Research Council, Washington, D.C., 1993, pp. 40–42.
31. Allen, J. G., H. S. Levinson, G. C. Garfield, and A. Coppoletta. South over the Alleys: The Rise, Fall and Rebuilding of Chicago's South Side 'L'. Presented at 88th Annual Meeting of the Transportation Research Board, Washington, D.C., Jan. 2009.
32. Allen, J. G., H. S. Levinson, B. G. Moffat, and G. C. Garfield. Chicago's Lake Street 'L': Ups, Downs, and a Rebound. Presented at 89th Annual Meeting of the Transportation Research Board of the National Academies, Washington, D.C., 2010.
33. Gash, S. M., G. C. Garfield, and D. C. Bollinger. A Decade of the Red Line and Green Line. *First & Fastest*, Summer 2003, p. 15.
34. Allen, J. G., and S. H. Abrams. Chicago's North Side 'L': The First Hundred Years. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 1793, Transportation Research Board of the National Academies, Washington, D.C., 2002, pp. 21–32.
35. Hinz, G. What's Wrong with the CTA? *Crain's Chicago Business*, Jan. 22, 2007.
36. *North Red and Purple Modernization Project. Environmental Impact Statement Scoping Information*. Chicago Transit Authority, Chicago, Ill., Jan. 2011.
37. *Derailment of Chicago Transit Authority Train Number 220 Between Clark/Lake and Grand/Milwaukee Stations, Chicago, Illinois, July 11, 2006*. National Transportation Safety Board, Washington, D.C., Sept. 11, 2007.
38. Wolinsky, J. Putting the Rapid Back into Transit. *Metro Report International*, March 2010, p. 20.
39. Effective Oversight of RTA Track Inspection Processes—A Case Study of the Chicago Transit Authority. *Rail Transit Safety Quarterly Newsletter* (U.S. Department of Transportation), summer 2009, pp. 8–13.
40. *2007—The Year of Decision: Regional Transportation Strategic Plan*. Regional Transportation Authority, Chicago, Ill., Feb. 8, 2007.
41. *Regional Transportation Authority Capital Asset Condition Assessment*. Prepared for the Regional Transportation Authority, Chicago, Ill., Aug. 2010.
42. *Transit State of Good Repair: Beginning the Dialogue*. Federal Transit Administration, Oct. 2008.
43. *Rail Modernization Study: Report to Congress*. Federal Transit Administration, April 2009, p. 2.
44. *2010 National State of Good Repair Assessment*. Federal Transit Administration, June 2010.
45. Gallucci, G., and J. G. Allen. Regional Transit Performance Measures at Chicago's Regional Transportation Authority. Presented at 90th Annual Meeting of the Transportation Research Board, Washington, D.C., 2010.

*The views are those of the authors and do not necessarily reflect the official policy of any organization.*

*The Rail Transit Infrastructure Committee peer-reviewed this paper.*

# Development of Base Train Equivalents to Standardize Trains for Capacity Analysis

Yung-Cheng (Rex) Lai, Yun-Hsuan Liu, and Tzu-Ya Lin

A conventional railway system usually has multiple train types with various service patterns operating on the same line. Differences in train characteristics lead to varied capacity effects on the system. "Rail line capacity" is commonly defined as the maximum number of trains that can be operated on a section of track with an expected level of service within a given time period. However, a particular unit (trains/hour or trains/day) does not reflect the train type the unit refers to. In this study, a new concept is proposed, namely, the base train equivalent (BTE), along with a standardization process to classify different train types in accordance with the particular type defined by the user. This concept is similar to the passenger car equivalent, which converts trucks to passenger car units in classifying highway transportation. A delay-based approach is also developed to determine BTEs on the basis of results obtained from two common capacity evaluation methods: parametric capacity analysis and simulation. With the proposed method, capacity measurements from different lines or systems can be compared and evaluated, resulting in meaningful and useful attributes.

With increasing traffic congestion and concerns about energy constraints and greenhouse gas emissions, considerable attention has been given to the increasing potential of intercity freight and passenger railway systems as productive means to resolve these nationwide problems. Identifying the magnitude and type of needed capacity improvements to accommodate a desired type of service is crucial in assisting public and private financing of capacity investments.

In a conventional railway system, trains with different characteristics commonly operate on the same track. Differences in train characteristics lead to different capacity effects on the system (1–4). Rail line capacity is commonly defined as the maximum number of trains that can be operated on a section of track with an expected level of service within a given time period (5–8). A particular unit (trains/hour or trains/day) does not reflect the train type the unit refers to. For example, 30 trains per day can refer to 30 passenger trains or freight trains per day; it may even jointly refer to a mix of passengers and freight trains. Thus, the impacts of their capacity differ significantly. To allow for accurate capacity calculation, the classification of different train types must be converted into a standard unit.

For highway capacity analysis, all types of vehicles are treated as passenger car units (PCUs) in the passenger car equivalent (PCE) approach. Various methods have been implemented to assess PCE for highway capacity analysis, including volume-to-capacity (V/C)

ratio, speed and density, headway, and delay approaches. The V/C ratio compares different traffic conditions with the same V/C ratio to determine PCE (9–11). Because the level of service is defined as the average running speed and density in the *Highway Capacity Manual*, some studies followed the change and computed PCE on the basis of equal speed or density method (12–14). Headway-based methods have also been proposed to compute PCE if headway between adjacent cars is available or can be computed (15–17). The model proposed by Krammes and Crowley derives the heavy vehicle adjustment factor (fHV) from the *Highway Capacity Manual* (18). Meanwhile, the delay-based approach has been developed for better estimation of the effects of heavy vehicles (19–21). The appropriate method for each type of capacity model should be consistent with the model characteristics.

In the railway domain, past studies identified the impact of heterogeneity on capacity (1, 2, 22, 23); however, there is no standard mechanism to evaluate the effect of heterogeneity and then convert different types of trains into a standard unit. In the current study, a new concept called base train equivalent (BTE) is proposed, along with a standardization process to classify different types of trains to a particular train type (base train unit, or BTU) as defined by the user. This concept is similar to PCE, which converts trucks to PCUs in classifying highway transportation. An approach is also developed to determine BTE on the basis of results obtained from two common capacity evaluation methods: parametric capacity analysis and simulation (8, 24–26). With the proposed method, the unit of rail capacity can be standardized, the impact of an additional train can be easily assessed, and capacity measurements from different lines or systems can be compared and evaluated, resulting in meaningful and useful attributes.

## ESTIMATION OF BTE IN RAIL LINE CAPACITY ANALYSIS

Similar to PCE, BTE is an attribute used to convert all train types to a standardized train unit or BTU. BTE can be defined as the ratio of the effects of a non-base train to a base train. Each railway system has various train types and services and may have different classifications of the base train. Therefore, instead of assigning a particular train type as BTU, the current study aims to develop a general concept of BTE and BTU to allow users to select their own BTUs.

Numerous approaches and tools have been developed to determine rail line capacity. These approaches can be categorized into the following three groups: (a) theoretical, (b) detailed simulation, and (c) parametric (23, 27, 28). In the theoretical approach, mathematical formulas or algebraic expressions are generally developed according to railway infrastructure to determine railway capacity (23). Simulation models estimate delay or capacity on the basis of given infrastructure configurations and decision rules of train dispatchers (22, 23, 29, 30). Parametric capacity models fill the gap between detailed simulation and simple theoretical formulas by focusing on

Department of Civil Engineering, National Taiwan University, Room 313, Civil Engineering Building, No. 1, Roosevelt Road, Section 4, Taipei 10617, Taiwan. Corresponding author: Y.-C. Lai, yclai@ntu.edu.tw.

*Transportation Research Record: Journal of the Transportation Research Board*, No. 2289, Transportation Research Board of the National Academies, Washington, D.C., 2012, pp. 119–125.  
DOI: 10.3141/2289-16



key elements of line capacity so as to quickly highlight “bottlenecks” in the system (8, 31, 32).

The primary tool used by all North American Class 1 railroads for capacity analysis is the Rail Traffic Controller (RTC) from Berkeley Simulation Software (28, 33, 34). Aside from simulation models, the parametric model developed by the Railway Canadian National (CN) is also widely known for strategic capacity planning (8, 35). The present study aims to develop an approach to determine BTEs in the capacity analysis results obtained from two common capacity evaluation tools.

As mentioned earlier, the appropriate method to estimate BTE should be consistent with the characteristics of the selected rail line capacity model. Because delay is the output of capacity analysis based on parametric and simulation models, a delay-based method is established to determine BTEs for the standardization process. In the following sections, the delay-based BTE method is introduced. The following section describes how it is used in parametric and simulation analysis.

### Delay-Based BTE Model

In highway capacity analysis, the delay-based approach is used to better estimate the effects of heavy vehicles (19, 20, 21). Unlike headway-based methods that consider only the excess headway consumed by trucks, the delay-based approach fully considers the delay in the traffic stream caused by heavy vehicles (19). The same concept was adopted in this study to develop the delay-based BTE equation for rail line capacity analysis. This equation can be written as follows:

$$E^D = 1 + \frac{\Delta d_i}{d_b} \quad (1)$$

where

$E^D$  = delay-based BTE,

$\Delta d_i$  = additional delay caused by one non-base train in mixed flow, and

$d_b$  = delay of one base train in base flow.

The computation process requires two different traffic flows (i.e., base and mixed flows) obtained from the capacity analysis. Base flow consists of only base trains, whereas mixed flow consists of both non-base and base trains. To determine the BTE, Equation 1 is then used to estimate the effect of one non-base train compared with one base train.

The additional delay of one non-base train can be obtained by subtracting the delay of the base flow from the delay of the mixed flow and then dividing the value by the number of non-base trains. The delay of one base train refers to the quotient of base flow delay and the total number of base trains. If the fractions are reduced to a common denominator, Equation 1 can be reformulated as

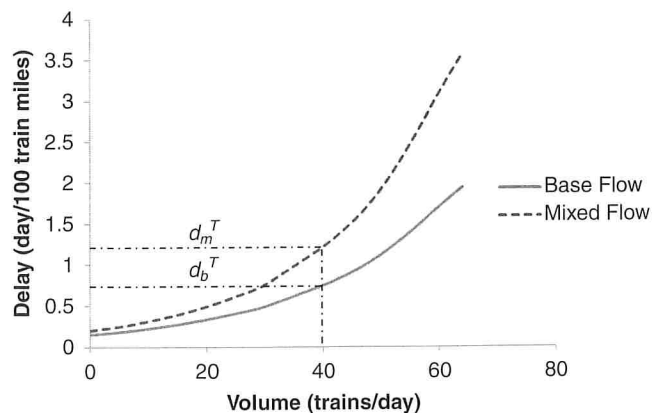


FIGURE 1 Delay-volume relationship of base and mixed flows.

$$E^D = \frac{d_b + \Delta d_i}{d_b} = \frac{\frac{d_b^T}{N} + \frac{(d_m^T - d_b^T)}{(N \times P)}}{\frac{d_b^T}{N}} = \frac{d_b^T + (d_m^T - d_b^T)P}{d_b^T} \quad (2)$$

where

$d_b^T$  = delay of total trains in base flow,

$d_m^T$  = delay of total trains in mixed flow,

$N$  = total number of trains, and

$P$  = percentage of non-base trains.

Equation 2 shows that with the same number of trains, BTE can be directly calculated on the basis of the total delay in the base and mixed flows. Figure 1 is an illustration of the delay-volume curves of the base and mixed flows. With the same number of trains (e.g., 40 trains/day), BTE can be obtained by deriving  $d_b^T$ ,  $d_m^T$ , and the number of base and non-base trains according to Equation 2.

Because the BTE model is based on delay obtained from the capacity analysis, the value of BTE would differ with changes in key capacity factors that include several route and operating parameters (8, 28, 35). For a particular network, all possible BTEs according to network characteristics are required to convert traffic volume into BTUs. Therefore, an implementation process of the BTE model was designed to obtain necessary BTEs.

### Implementation Process of Delay-Based BTE Model

Figure 2 shows the implementation process of the delay-based BTE model in computing possible BTEs. On the basis of a set of possible

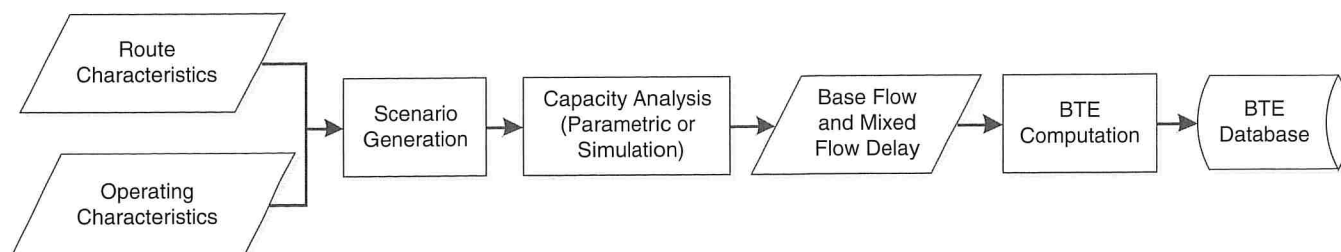


FIGURE 2 Implementation process of delay-based BTE model.

TABLE 1 Part of BTE Database

Volume	Percentage of Non-Base Trains	Speed Ratio	SDS (mi)	SGS (mi)	BTE
16	25	1.40	11.0	1.375	2.16
16	25	1.40	11.0	2.750	2.19
16	25	2.33	5.5	1.375	3.94
16	25	2.33	5.5	2.750	3.98
16	50	1.40	11.0	1.375	1.96
16	50	1.40	11.0	2.750	1.99
16	50	2.33	5.5	1.375	3.16
16	50	2.33	5.5	2.750	3.21
24	25	1.40	11.0	1.375	2.17
24	25	1.40	11.0	2.750	2.20
24	25	2.33	5.5	1.375	3.95
24	25	2.33	5.5	2.750	3.99
24	50	1.40	11.0	1.375	1.98
24	50	1.40	11.0	2.750	2.00
24	50	2.33	5.5	1.375	3.19
24	50	2.33	5.5	2.750	3.23

route and operating characteristics, the process first generates possible scenarios and then performs capacity analysis for each of the scenarios by using parametric or simulation models. All delays from the mixed and base flows in the capacity analysis are used to compute possible BTEs. Finally, these results are stored in the BTE database. Table 1 shows a part of the BTE database sorted into five capacity factors: volume, heterogeneity, speed, siding spacing (SDS), and signal spacing (SGS).

With the data in the BTE database, BTU conversion was conducted by using the process illustrated in Figure 3. According to specific route and traffic characteristics of a link in the network (such as a subdivision), volume, and composition, the appropriate BTE in the database can be chosen. The calculation of BTU then takes the BTE along with volume and composition to obtain the result. By repeating this process for every link in the network, different lines can thus be compared and evaluated according to the same unit—BTU.

## CASE STUDY

To compute for BTE values, a case study was first conducted during the implementation process for both parametric capacity analysis and simulation. The case study was based on a set of inputs representing the typical characteristics of a midwestern North American

TABLE 2 Possible Route and Operating Characteristics Considered

Parameter	Levels	Number of Levels
Siding spacing (mi)	5.5, 11, 16.5	3
Signal spacing (mi)	1.375, 2.75, 5.5	3
Volume (trains per day)	16, 24, 32	3
Heterogeneity (% of coal trains)	0, 12.5, 25, 50, 75, 87.5, 100	7
Speed ratio (mph/mph)	70/30, 70/50	2

NOTE: Total scenarios examined = 378.

single-track main line. A sensitivity analysis was also conducted to evaluate the relevance of each capacity factor in the BTE. Finally, the BTE values obtained from the parametric capacity analysis were used in a corridor capacity analysis to emphasize the importance of heterogeneity in capacity computation.

Five key capacity factors with several levels were selected for the implementation process of this case study according to research done by Krueger (8) and Dingler (36). These factors included SDS, SGS, volume, heterogeneity, and speed ratio. The possible levels in each factor are shown in Table 2. A total of 378 scenarios were examined, and each had corresponding BTE values. The current study assumed that there were only two types of trains: (a) intermodal trains as the base trains and (b) coal trains as the non-base trains (Table 3). Moreover, heterogeneity was defined as the percentage of coal trains (% of coal trains) in the total number of trains.

## BTE Computation Based on Parametric Capacity Analysis

The CN parametric model was used to perform the capacity analysis for all 378 scenarios (8). BTEs for all scenarios were obtained and stored in the database. Sensitivity analysis was then performed to evaluate the relevance of each capacity factor in the BTE (Figure 4). Because it was impossible to reveal all the BTEs, part of the results are shown in Figure 4 along with the sensitivity analysis. According to Figure 4, *a–d*, BTE values generally range from 2 to 6. Results of the sensitivity analysis indicated that two out of five capacity factors (i.e., heterogeneity and speed ratio) were quite sensitive to BTE. These results are intuitive because differences in heterogeneity and speed ratio are directly related to the differences in train characteristics.

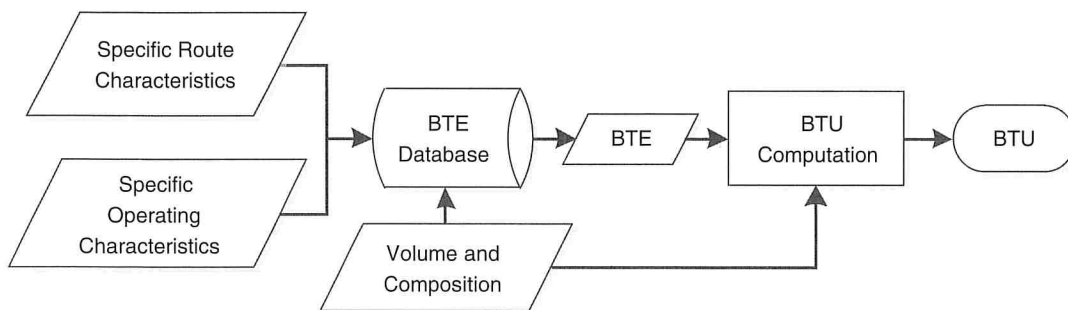


FIGURE 3 Process diagrams of BTU conversion according to BTEs.

**TABLE 3** Characteristics of Intermodal and Coal Trains

Characteristic	Intermodal	Coal
Departure load (units)	93	115
Loading weight (tons)	5,900	16,445
Train length (ft)	5,649	6,325
Locomotives (hp)	5 × 4,300	3 × 4,300
Maximum speed (mph)	70	50

NOTE: hp = horsepower.

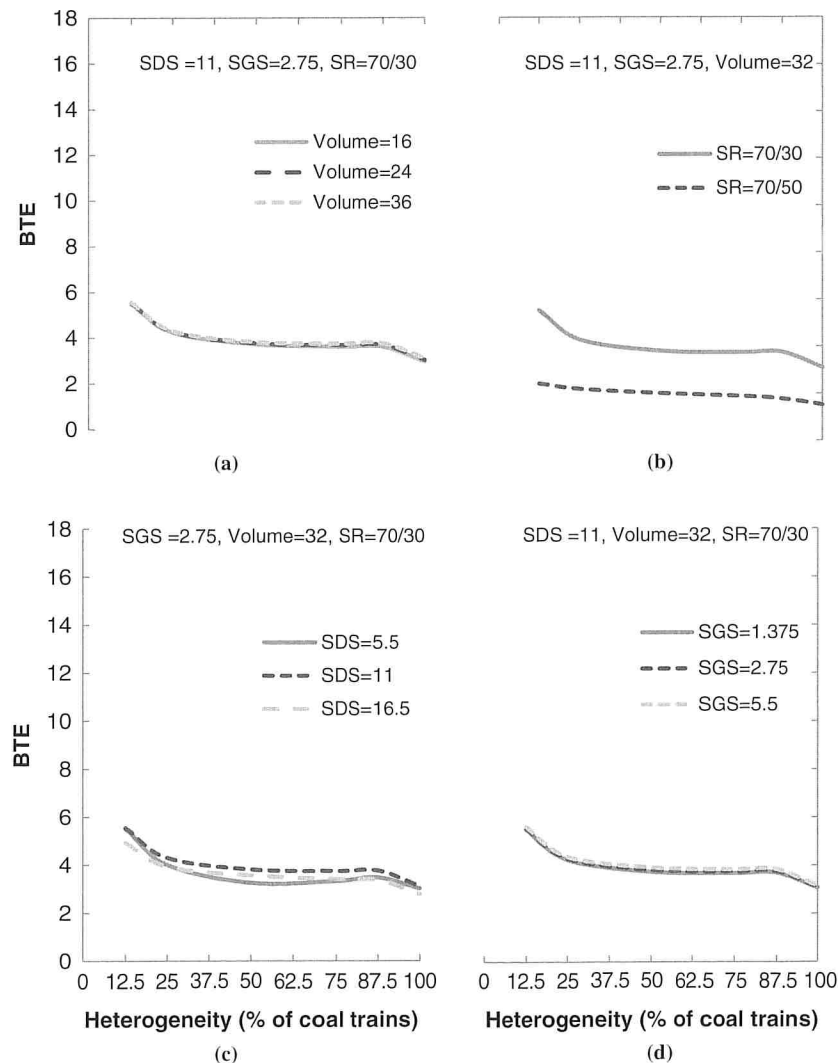
### BTE Computation Based on Simulation Analysis

RTC was used to perform the simulation analysis. For each scenario, the simulation was conducted under 30 different random seeds, after which the average value was obtained to eliminate the stochastic variance. One random seed had its own exclusive departure time, which can vary by 15 min before or after the scheduled departure time. The purpose of adding random

seeds was to simulate real-time train operation as much as possible. The number of random seeds was determined by a statistical method.

BTE results obtained through RTC simulations were quite different from those of the parametric analysis. Because of randomness, the simulation results were not as uniform as those from the parametric capacity analysis. As shown in Figure 5, the BTE values range from 2 to 16 depending on the scenarios. By contrast to SGS, heterogeneity, volume, speed ratio, and SDS were all found to be sensitive to BTE. The peak value of BTE at a certain volume occurred at 12.5% or 25% of coal trains (Figure 5a), similar to the results observed in the parametric capacity analysis. In Figure 5a, BTE values for 16 and 24 trains per day were quite similar, and the BTE jumped to a much higher level at 32 trains per day, indicating that the capacity bottleneck was reached (Figure 5a). SDS became a factor sensitive to BTE according to simulation results because the delay grew rapidly as SDS increased.

To examine the outcome of the CN parametric model and RTC simulations, delay–volume curves based on different tools were plotted on the same graph (Figure 6), which indicated that results of the tools were quite different. Therefore, BTE values were not con-



**FIGURE 4** Relationships between BTE, heterogeneity, and (a) volume, (b) speed ratio, (c) siding spacing, and (d) signal spacing, on basis of parametric analysis.

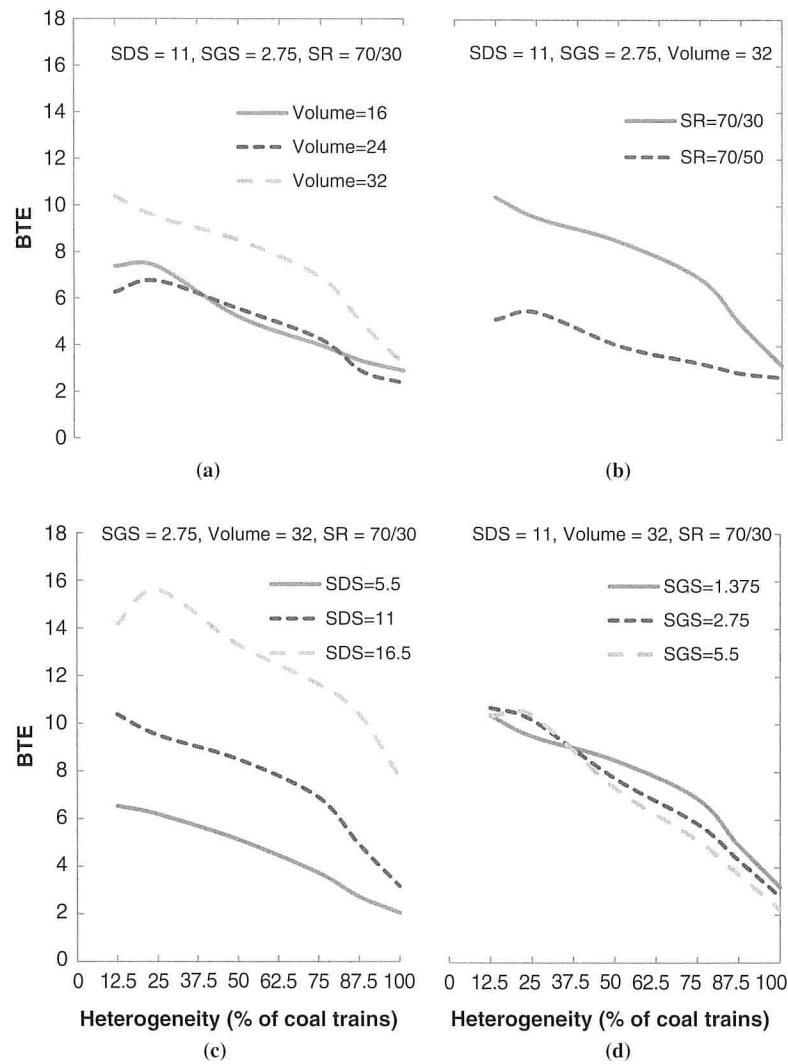


FIGURE 5 Relationships between BTE, heterogeneity, and (a) volume, (b) speed ratio, (c) siding spacing, and (d) signal spacing, on basis of simulation analysis.

sistent between these two capacity evaluation tools. The CN parametric model was developed to perform capacity analysis within a particular range of parameters. The effect from traffic mix is also less significant in the CN parametric model as opposed to in RTC simulations. Therefore, it is important to use appropriate BTEs according to the analytical tools for capacity analysis.

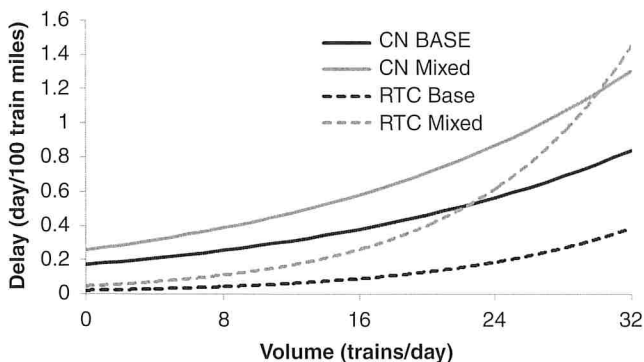


FIGURE 6 Delay-volume curves of the two models (SDS = 11, SGS = 2.75, with 50% coal trains).

### Application of BTU Conversion Process in Capacity Analysis

In this section, the BTU conversion process (Figure 3) was implemented on a hypothetical corridor with seven subdivisions on the basis of the BTEs obtained from the parametric capacity analysis. The basic characteristics of this corridor are listed in Table 4. Without the conversion, the traffic volume for every subdivision was the same at 32 trains per day; however, these 32 trains had different traffic compositions (i.e., heterogeneity), thus resulting in different capacity effects. With the BTU conversion, the subdivision with the most traffic was clearly identified, that is, Subdivision D with 75 BTU. By applying the concept of BTE, the capacity in different subdivisions or systems can be compared on the basis of a consistent standard.

### FURTHER DISCUSSION

For years, highway transportation capacity planners have benefited from the *Highway Capacity Manual*, which offers a well-established standard methodology developed over many years and has been widely used in the highway sector. Unfortunately, no such model exists for

TABLE 4 Basic Parameters and Results of BTE Analysis

Subdivision	SDS (mi)	SGS (mi)	No. of Base Trains	No. of Non-Base Trains	Speed Ratio	BTE	Volume per Day	BTU per Day
A	11	2.75	32	0	NA	NA	32	32
B	11	1.375	28	4	70/30	5.56	32	50
C	5.5	2.75	24	8	70/50	1.82	32	39
D	16.5	5.5	16	16	70/30	3.66	32	75
E	11	5.5	28	4	70/50	2.24	32	37
F	5.5	1.375	24	8	70/30	3.93	32	55
G	16.5	1.375	16	16	70/50	1.91	32	47

NOTE: No. = number; NA = not available.

railroad transportation, and current railroad capacity planning requires the use of intensive simulations. Aside from its inefficiency for railroad planners, difficulty also manifests in dealing with various parties who wish to implement or expand passenger rail operations on freight railroads or criticize freight railroads for current operations. There is a need for an objective, realistic tool that can be used to establish realistic expectations for freight railroad performance and also to provide an assessment of the disproportionate effect of passenger rail operation on near- or at-capacity freight rail lines.

The creation of a standard rail line capacity model requires the development of a method by which the consumed capacity and maximum capacity of a route can be calculated. Both are typically measured in rail systems every day; unfortunately, just as in highway transportation, different types of trains use the capacity differently. To allow capacity calculation, these different train types must be converted into a standard unit. In this study, a new concept is proposed, the BTE, along with a standardization process to classify different train types in accordance with the particular type defined by the user. With the proposed method, the unit of rail capacity can be standardized as BTU for a period of time. This results in a more meaningful attribute compared with traffic volume as shown in the case study (Table 4). This novel concept makes it possible to compare and evaluate the capacity across different subdivisions, lines, or systems. In addition, the effect of adding an additional train to a particular scenario can be easily captured because BTE represents the ratio of the effects of a non-base train to the effects of a base train.

The case study also demonstrates the importance of using appropriate BTEs according to the selected capacity analysis tool, that is, parametric capacity analysis or simulation. If railway agencies and companies can form a consensus on the standard unit along with possible route and train characteristics, a comprehensive BTE database with both parametric and simulation results can be developed (Figure 2). This database can be a significant accomplishment, allowing the railway capacity manual to standardize capacity analysis.

## CONCLUSION

This study proposes BTEs and a standardization process to convert different train types to a particular train type as defined by users. A delay-based model has also been developed to determine BTEs in the results obtained from two capacity evaluation methods: parametric capacity analysis and simulation. With the proposed method, the unit of rail capacity can be standardized, the effect of an additional train can be easily assessed, and capacity measurements from different lines or systems can be compared and evaluated, thus resulting in meaningful and useful attributes.

## ACKNOWLEDGMENTS

The authors are grateful to Eric Wilson and Mark Dangler for their assistance in this research. This project was funded by the National Science Council (NSC) of Taiwan.

## REFERENCES

1. Dangler, M. H., Y.-C. Lai, and C. P. L. Barkan. Impact of Train Type Heterogeneity on Single-Track Railway Capacity. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 2117, Transportation Research Board of the National Academies, Washington, D.C., 2009, pp. 41–49.
2. Harrod, S. Capacity Factors of a Mixed Speed Railway Network. *Transportation Research Part E*, Vol. 45, 2009, pp. 830–841.
3. Lai, Y.-C., M. H. Dangler, C.-E. Hsu, and P.-C. Chiang. Optimizing Train Network Routing with Heterogeneous Traffic. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 2159, Transportation Research Board of the National Academies, Washington, D.C., 2010, pp. 69–76.
4. Murali, P., M. M. Dessouky, F. Ordonex, and K. Palmer. A Delay Estimation Technique for Single and Double-Track Railroads. *Transportation Research Part E*, Vol. 46, 2009, pp. 483–495.
5. Kittelson and Associates, Inc. *TCRP Report 100: Transit Capacity and Quality of Service Manual*, 2nd ed. Transportation Research Board of the National Academies, Washington, D.C., 2003.
6. Kozan, E., and R. L. Burdett. A Railway Capacity Determination Model and Rail Access Charging Methodologies. *Transportation Planning and Technology*, Vol. 28, No. 1, 2005, pp. 27–45.
7. Kozan, E., and R. L. Burdett. Techniques for Absolute Capacity Determination in Railways. *Transportation Research Part B*, Vol. 40, 2006, pp. 616–632.
8. Krueger, H. Parametric Modeling in Rail Capacity Planning. *Proc., Winter Simulation Conference*, Phoenix, Ariz., 1999.
9. Ingle, A. *Development of Passenger Car Equivalents for Basic Freeway Segments*. MS thesis. Virginia Polytechnic Institute and State University, Blacksburg, 2004.
10. Linzer, E. M., R. P. Roess, and W. R. McShane. Effect of Trucks, Buses, and Recreational Vehicles on Freeway Capacity and Service Volume. In *Transportation Research Record 699*, TRB, National Research Council, Washington, D.C., 1979, pp. 17–24.
11. May, A. *Traffic Flow Fundamentals*. Prentice-Hall, Englewood Cliffs, N.J., 1990, pp. 247–253.
12. Greenshields, B. D. A Study of Traffic Capacity. *Proceedings of the Highway Research Board*, Washington, D.C., 1934.
13. Huber, M. J. Estimation of Passenger-Car Equivalents of Trucks in Traffic Stream. In *Transportation Research Record 869*, TRB, National Research Council, Washington, D.C., 1982, pp. 60–70.
14. Van Aerde, M., and S. Yagar. Capacity, Speed, and Platooning Vehicle Equivalents for Two-Lane Rural Highways. In *Transportation Research Record 971*, TRB, National Research Council, Washington, D.C., 1984, pp. 58–67.
15. Krammes, R. A., and K. W. Crowley. Passenger Car Equivalents for Trucks on Level Freeway Segments. In *Transportation Research Record 1091*, TRB, National Research Council, Washington, D.C., 1986, pp. 10–17.



16. Seguin, E., K. Crowley, and W. Zweig. *Passenger Car Equivalents on Urban Freeways*. Report DTFH61-80-C-00106. FHWA, U.S. Department of Transportation, 1982.
17. Werner, A., and J. F. Morrall. Passenger Car Equivalencies of Trucks, Buses, and Recreational Vehicles for Two-Lane Rural Highways. In *Transportation Research Record 615*, TRB, National Research Council, Washington, D.C., 1976, pp. 10–17.
18. *Special Report 209: Highway Capacity Manual*, 3rd ed. TRB, National Research Council, Washington, D.C., 1985.
19. Benekohal, R. F., and W. Zhao. Delay-based Passenger Car Equivalents for Trucks at Signalized Intersections. *Transportation Research Part A*, Vol. 34, 1999, pp. 437–457.
20. Cunagin, W. D., and C. J. Messer. Passenger Car Equivalents for Rural Highways. In *Transportation Research Record 905*, TRB, National Research Council, Washington, D.C., 1983, pp. 61–68.
21. Rodriguez-Seda, J. D., and R. F. Benekohal. Methodology for Delay-Based Passenger Car Equivalencies for Urban Transit Buses. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 1988, Transportation Research Board of the National Academies, Washington, D.C., 2006, pp. 127–137.
22. Bronzini, M. S., and D. B. Clarke. Estimating Rail Line Capacity and Delay by Computer Simulation. *Transportation Forum*, Vol. 2, No. 1, 1985, pp. 5–11.
23. Abril, M., F. Barber, L. Ingolotti, M. A. Salido, P. Tormos, and A. Lova. An Assessment of Railway Capacity. *Transportation Research Part E*, Vol. 44, No. 5, 2008, pp. 774–806.
24. Wilson, E. *Rail Traffic Controller (RTC) Brochure*. Berkeley Simulation Software, Berkeley, Calif., 2008.
25. Thompson, W. CREATE Update. *Proc., AREMA 2006 Annual Conference*, Landover, Md., 2006.
26. Lai, Y.-C., and C. P. L. Barkan. A Comprehensive Decision Support Framework for Strategic Railway Capacity Planning. *ASCE Journal of Transportation Engineering*, Vol. 137, No. 10, 2011, pp. 738–749.
27. Gorman, M. H. Statistical Estimation of Railroad Congestion Delay. *Transportation Research Part E*, Vol. 45, 2009, pp. 446–456.
28. Lai, Y.-C. *Increasing Railway Efficiency and Capacity Through Improved Operations, Control and Planning*. PhD dissertation. University of Illinois, Urbana, 2008.
29. Fransoo, C., and J. Bertranda. Aggregate Capacity Estimation Model for the Evaluation of Railroad Passing Constructions. *Transportation Research Part A*, Vol. 34, 2000, pp. 35–49.
30. Vromans, M. J. C. M., R. Dekker, and L. G. Kroon. Reliability and Heterogeneity of Railway Services. *European Journal of Operational Research*, No. 172, 2006, pp. 647–655.
31. Lai, Y.-C., and C. P. L. Barkan. Enhanced Parametric Railway Capacity Evaluation Tool. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 2117, Transportation Research Board of the National Academies, Washington, D.C., 2009, pp. 33–40.
32. Prokopy, J. C., and R. B. Rubin. *Parametric Analysis of Railway Line Capacity*. DOTFR-5014-2. Federal Railroad Association, U.S. Department of Transportation, 1975.
33. *The Long Term Financial Feasibility of the Northwestern Pacific Railroad*. Draft final. Parsons Brinckerhoff Quade & Douglas, Inc., 2002.
34. Washington Group International, Inc. *RTC Simulations—LOSSAN North Railroad Capacity and Performance Analysis*. LOSSAN Rail Corridor Agency and IBI Group, 2007.
35. Vantuono, W. C. Capacity Is Where You Find It: How BNSF Balances Infrastructure and Operations. *Railway Age*, Feb. 1, 2005.
36. Dingler, M. *The Impact of Operational Strategies and New Technologies on Railroad Capacity*. MS thesis. University of Illinois at Urbana-Champaign, 2011.

---

*The Freight Rail Transportation Committee peer-reviewed this paper.*

# Methodological Framework for Analyzing Ability of Freight Rail Customers to Forecast Short-Term Volumes Accurately

Stephan Moll, Ulrich Weidmann, and Andrew Nash

The freight transport business is extremely challenging for railways because transport by truck has intrinsic advantages in flexibility and quality. Providing freight customers with flexible scheduling is particularly difficult because optimizing an interconnected rail operating plan is more difficult than arranging for shipment by truck. In this environment it would be helpful if shippers could provide railways with accurate demand forecasts. However, the ability to forecast rail freight transport differs strongly by shipper and commodity type. The goal of this research is to develop a methodological framework to understand better the characteristics that influence the ability of freight shippers to prepare accurate forecasts of rail demand. This information will help railways increase productivity by improving their ability to develop optimized schedules. It will help railways decide when to rely on shipper forecasts and provide a benchmark for identifying shippers that can provide accurate forecasts. The paper describes the methodological framework and presents results from a case study application to illustrate the practical applicability of the proposed framework.

The freight transport business is extremely competitive. Freight transport is particularly challenging for railways because transport by truck provides high flexibility for customers. As a result railways are faced with a difficult balancing act between providing flexible customer solutions and operating at high productivity—in addition to strong pricing pressure.

The requirement for providing customers with flexible scheduling is particularly difficult for railways because numerous infrastructural constraints make the optimization of rail operations complicated. Nevertheless, shipping companies can order unit trains on a weekly basis in Switzerland. This weekly ordering procedure leaves the freight railway with only a few days for creating timetables and duty schedules.

In this environment it would be very helpful if freight shippers could provide railways with accurate demand forecasts so the railways could optimize resource planning. However, the ability to forecast rail freight transport differs strongly among companies and commodity types.

The goal of this research is to develop a methodological framework to understand better the characteristics that influ-

ence the ability of freight shippers to prepare accurate forecasts of rail demand. Understanding the quality of company-provided rail demand forecasts can help railways increase productivity by improving their ability to develop optimized schedules. Furthermore, the results can be used as a benchmark to identify shipping companies that have the ability to provide accurate forecasts. Railways can use this benchmarking information to request forecasts from companies with the potential for providing accurate forecasts and can adjust service provisions for companies that have lower potential for providing accurate forecasts.

The next section describes the approach used for developing the methodological framework. A detailed description of the framework and the framework characteristics is presented next. A specific application of the framework is then described, followed by conclusions.

## METHODOLOGY AND DATA COLLECTION

The proposed framework for analyzing and describing rail transport planning at shipping companies is based on the concept of morphological analysis. Fritz Zwicky developed morphological analysis in the 1940s as a method “for structuring and investigating the total set of relationships contained in multi-dimensional, non-quantifiable, problem complexes” (1).

The basic idea of morphological analysis is to break a subject down into a set of fundamental features that describe the subject as completely as necessary to solve a specific problem (2). Next, the possible values for each feature are identified. This approach ensures that morphological schemes show the full range of possibilities and that no possibility is neglected. This full range enables planners to complete a structured and comprehensible analysis of complex problems.

Morphological analysis has been applied to many diverse subjects. In the field of logistics morphological schemes have been created both to describe in-house logistics planning (3) and to characterize supply chains among manufacturing companies (4). The goal of this research is to create a morphological scheme focusing on freight rail transport planning.

The basic structure of the proposed morphological scheme was derived from the supply chain planning matrix (5). An adapted version of this matrix is illustrated in Figure 1.

Three main categories from the matrix are especially relevant to rail freight transport planning:

- Transport logistics organization—defined as the infrastructure and organizational limits for transport planning; typical

S. Moll and U. Weidmann, Swiss Federal Institute of Technology (ETH), Institute for Transport Planning and Systems (IVT), Wolfgang-Pauli-Strasse 15, Zurich 8093, Switzerland. A. Nash, Vienna Transportation Strategies, Bandgasse 21/15, Vienna 1070, Austria. Corresponding author: S. Moll, moll@ivt.baug.ethz.ch.

*Transportation Research Record: Journal of the Transportation Research Board*, No. 2289, Transportation Research Board of the National Academies, Washington, D.C., 2012, pp. 126–133.  
DOI: 10.3141/2289-17

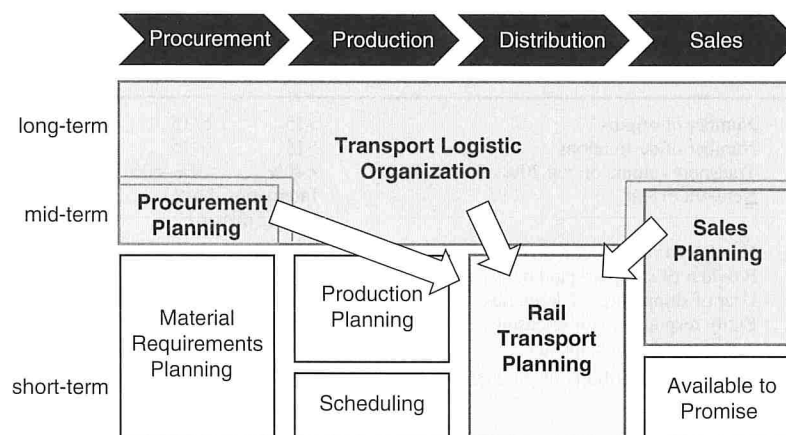


FIGURE 1 Main areas relevant to transport planning.

examples are inventory capacities and the geographical distribution of goods being shipped;

- Sales and procurement planning—essentially the precursors to transport demand; vague sales and procurement estimates significantly affect the quality of transport forecasts (the choice of whether sales or procurement forecasts need to be considered depends on who is responsible for making the shipping decisions, the supplier or the receiver); and
- Rail transport planning—this includes many factors ranging from transport flexibility to the form and quality of forecasts and orders.

In this research the most relevant features in these three categories were identified on the basis of expert interviews and literature review. The literature provided important clues on constraints influencing transport logistics, including the production process, stock capacities, and transport route (6, 7). A series of in-depth personal interviews with transport experts from a large Swiss freight railway and eight representative shippers was the main source of information. The companies are located in Switzerland and operate in very different sectors, including mineral oil, gravel, and iron and steel.

The next section describes the proposed morphological framework and characteristics in more detail.

## PROPOSED MORPHOLOGICAL FRAMEWORK

The proposed morphological framework is organized in relation to category, subcategory, feature, and values. The research goal was to identify the feature values that are “more likely” to lead to accurate shipping forecasts by the shippers. Railways could then use this information to determine the ability of shippers to prepare accurate forecasts and to help determine customer specific service strategies (e.g., pricing and service frequency). This information, in turn, can be used to help optimize railway schedule development.

In the tables, the values for each feature are arranged so that the railway’s expected probability for receiving an accurate forecast from the shipper increases from left to right.

### Transport Logistics Organization

There are four key elements of a transport logistics organization that affect rail freight planning: transport network structure, transport

responsibility, transport origin, and transport destination. These elements and their features are outlined below and summarized in Table 1.

### Transport Network Structure

The structure of the transport network has a fundamental effect on transport planning; put simply, the more complex the network, the more challenging is the planning. Presented in Table 1, the four transport planning features identified in this research for use in characterizing a transport network are

- Number of origins,
- Number of destinations,
- Transport volume of top 20% of routes, and
- Network extent.

The simplest transport network consists of one origin and one destination. Network complexity increases with the number of origins and destinations. Therefore the best situation for making forecasts is instances in which there is only one possible origin and one possible destination (shown by the value 1 being placed at the right-hand side of the table).

Planning complexity of networks is reduced if transports are concentrated mainly on a few important routes. This concentration can be measured by the distribution of transport volumes of the transport routes of a specific network. The proposed feature is the transport volume of the top 20% of routes, based on the number of shipped wagons. The higher this percentage value, the more concentrated are transport volumes on a few important transport routes of a transport network.

A network consisting of long-distance or international transport relationships or both is generally more challenging for transport planning because the chance of delays increases with longer distances and border crossings. This challenge is characterized by the feature “network extent.”

### Transport Responsibility

As shown in Table 1, transport responsibility consists of four main features affecting transport planning. The number of involved railways

TABLE 1 Transport Logistics Organization: Features and Proposed Values

Subcategory	Feature	Value			
Transport network structure	Number of origins	>15	6–15	2–5	1
	Number of destinations	>15	6–15	2–5	1
	Transport volume on top 20% of relations	<40%	40%–60%	60%–80%	>80%
	Network extent	International or long distance		Domestic or short distance	
Transport responsibility	Number of railways involved	Several		One	
	Provider of shipping-paid deliveries	Yes		No	
	User of shipping-paid deliveries	No		Yes, partly	Yes
	Entity responsible for releasing products being shipped	≠ freight payer		= freight payer	
Transport origin	Quantitative flexibility of production capacity at origin	Flexible in terms of time		Hardly flexible in terms of time	Not flexible in terms of time
	Inventory philosophy at origin	High inventory		Medium inventory	Minimum inventory (just in time)
	Entity in charge of origin site	≠ entity placing shipment order		= entity placing shipment order	
	Use of rail siding	Exclusive		Shared	
Transport destination	Quantitative flexibility of production capacity at destination	Flexible in terms of time		Hardly flexible in terms of time	Not flexible in terms of time
	Inventory philosophy at destination	High inventory		Medium inventory	Minimum inventory (just in time)
	Entity in charge of destination site	≠ entity placing shipment order		= entity placing shipment order	
	Use of rail siding	Exclusive		Shared	

is relevant mainly in international shipping in Europe. Although Europe's liberalized rail freight market allows railways to run trains in several countries, often this practice is not economically feasible and several railways are involved in completing the transport.

The more railways involved in the transport process, the more complex the planning because there is a high potential for delay when handing over trains (especially at national borders) and different railways have different punctuality standards. Furthermore, delays to one shipment may force companies to make short-term changes to other planned shipments to ensure adequate supply and distribution throughout the logistics chain; this necessity further increases transport planning complexity.

Shipping-paid deliveries are freight shipments in which the supplier pays and organizes the transport to the company ordering the product being shipped. In an ideal world the shipment would be forecast by the company ordering the product, but in reality many companies are not willing to do so. Besides, partly using the service of shipping-paid deliveries can be highly advantageous for companies if they bear responsibility for transports of basic demands and use shipping-paid deliveries for peak demands. This strategy has a stabilizing effect for users of shipping-paid deliveries in regard to transport planning, but can have the opposite effect for providers of shipping-paid deliveries. Their transport planning tends to become less stable and predictable.

Eventually if the entity responsible for releasing products being shipped is concurrently the entity ordering the transport service, then its ability to provide good forecasts to the freight railway is higher than if the entity ordering the service does not control release of the shipment.

### *Physical Characteristics at Transport Origin and Destination*

The third subcategory of a transport logistic organization is the physical situation at the transport origin and destination. As illustrated in Table 1 the same features and values are proposed for origins and destinations.

Transport origins and destinations can be either storage facilities or production sites. In both cases the ability to buffer demand for transport service either through storage facilities or as part of the production process influences the predictability of transport planning forecasts.

Quantitative flexibility of production capacity describes the temporal flexibility of production capacity (3). All other things being equal, a more flexible production capacity reduces the ability to forecast transportation demand precisely.

Inventory philosophy describes an aggregated set of storage depot characteristics from the perspective of transportation planning (because management of storage depots is a very complex optimization problem, the measure proposed in this research ignores aspects of storage depot management that do not directly affect transport planning). This research defines three levels of inventory philosophies:

- Minimum inventory. Storage depots with a minimum inventory are highly dependent on regular or accurately planned transport or both. Therefore the accuracy of forecasts from storage facilities with minimum inventory is generally very high because any deviations from transport schedules would risk a shortage or overflow of invento-

ries within days. Minimum inventory is typically found at companies following a Just-in-Time strategy, investing in logistics, production techniques, and training of personnel to minimize inventories.

- Medium inventory. Storage depots with a medium inventory have more transport flexibility than those with minimum inventory; therefore their transport forecasts tend to be less accurate (because, e.g., production does not need to be shipped away immediately).
- High inventory. Storage depots with a high inventory provide extreme transport planning flexibility because their high capacity means that they do not require regular supply or demand transport.

For production sites, transport demand forecasts are expected to be stable if production is not flexible in regard to time and there is a minimum inventory storage capacity available. In this case, production output must be shipped away immediately. Production sites with little quantitative flexibility are typically found in process industries (e.g., refineries).

For storage sites, the storage capacity is the only relevant feature. Transport demand forecasts are expected to be stable if there is little storage capacity.

The entity in charge of the transport site is another relevant feature. If the entity ordering service also controls the transport site, there is a higher chance of coordinated planning of production, inventory, and transport. This coordination helps ensure that requirements for achieving stable and accurate transport plans are more likely to be considered in inventory and production planning decisions and therefore forecasts should be accurate.

However, if the entity in charge of the transport site is not the entity providing the forecasts, there is a risk of unilateral inventory or production optimizations at the expense of transport forecasting accuracy. This risk applies especially if the entity in charge of the transport site is in a dominant position toward the entity responsible for ordering the transport service. Note that although generally the entity ordering service will be at either the origin or destination, there are also cases in which the same entity can be at both ends of the transport chain.

Shared use of a rail siding eventually forces companies to carefully plan their shipment loading and unloading time slots. This need for cooperation reduces short-term flexibility of transport planning, which means that forecasts coming from companies that share sidings should be more accurate.

## Sales and Procurement Planning

The second category of information is sales and procurement planning. In practice analysts consider either sales or procurement planning, depending on who is ordering the transport service. If the shipper is ordering service, sales planning is considered; and if the purchaser is ordering service, procurement planning is considered.

There are two subcategories: freight customer service characteristics and sales and procurement plan. Table 2 presents the features and values for sales and procurement planning.

The larger the shipment lot size, the more it affects rail transport planning. Large shipments are made with unit trains; modifying these orders often necessitates rescheduling shipper and carrier transport plans. Changes to smaller lot sizes—for example, single wagonload service—typically have a small effect on scheduling because of the large number of possibilities for adding a freight wagon to another train.

If delivery date flexibility is not provided, transport planning lacks an important option for optimization; however, transport forecasts become more stable because there is no possibility to modify delivery dates at short notice.

If delivery dates are flexible, it is possible to optimize scheduling by making short-term modifications; however, these short-term modifications do not necessarily increase the potential for increased rail transport system optimization. Quite the contrary, flexible delivery dates could be a sign that a shipper cannot determine an exact delivery date (e.g., goods being shipped are loaded but awaiting results of quality tests). In this case, transport planning at the railway would be reduced to a day-to-day activity with no dependable planning horizon.

Sales and procurement plans of shipping companies are the precursors to transport forecasts and therefore a major source of information for transport planning. The sales and procurement plan has three features: forecast time horizon, level of detail, and forecast technique. The value of these features is influenced by numerous factors including volatility of demand, flexibility of sales terms, and intensity of collaboration between customer and shipper.

The time horizon and level of detail values reflect the reasonably achievable accuracy in sales and procurement planning. As shown in Table 2, the longer the forecast period and the more precise the level of detail (i.e., a daily forecast rather than a monthly forecast), the better.

There are three basic demand estimation techniques: intuitive forecasts, mathematical forecasts, and firm orders. Sales or procurement forecasts developed with mathematical forecasting techniques [implying the existence of a quantifiable future demand behavior (3)], are generally more accurate than those made for irregular or sporadic demands (typically, intuitive forecasts). The highest accuracy is expected if planning can rely on firm orders only. No forecast would then be necessary.

An accurate and long-term sales and procurement plan generally leads to an accurate and long-term transport plan, as well. However, the reverse argument is not necessarily true. Transport planning can be accurate even if underlying sales and procurement plans are uncertain. However, in this case the shipping company must have many transport planning options available to manage demand deviations, which it can use to reduce the need for changes to the planned transport schedule (see subsection on transport flexibility).

TABLE 2 Sales and Procurement Planning: Features and Proposed Values

Subcategory	Feature	Value		
Freight customer service characteristics	Shipment lot size	Unit trains		Freight wagons
	Delivery date flexibility	±1 week or more	±1 day	Not flexible
Sales and procurement plan	Time horizon	<1 week	≥1 week	≥1 month
	Level of detail	Month	Week	Day
	Forecast technique	Intuitive	Mathematical	None (firm orders)



TABLE 3 Transport Planning: Features and Proposed Values

Subcategory	Feature	Value			
Transport plan	Time horizon	<1 week	≥1 week	≥1 month	
	Level of detail	Month	Week	Day	
	Revision cycle	Daily	Event driven	Weekly	No revision
Transport flexibility	Availability of transport alternatives at short notice	Yes		No	
	Latest possible time for placing transport order	≤1 day prior to shipping	≤1 week prior to shipping	>1 week prior to shipping	
	Transport vehicle ownership	Ad hoc leasing		Proprietary possession or long-term leasing	
	Nontransport measures for addressing demand deviations	None	Few	Many	
Forecast information provided to the freight railway	Forecast accuracy	<80%	80%–85%	85%–90%	>90%
	Share of shipments forecasted	<50%	50%–75%	75%–90%	>90%
	Forecast frequency	Irregular	Monthly	Weekly	Daily
	Forecast time horizon	≤1 week	2–3 weeks	≥1 month	
	Forecast level of detail	Month	Week	Day	
Order information	Order accuracy	<85%	85%–90%	90%–95%	>95%
	Dominant type of order change	Cancel shipment	Change date	Change destination	Additional goods to transport

## Transport Planning

The third category of information is transport planning. It considers the transport plan, transport flexibility, forecast information provided to the freight railway, and order information. These four subcategories and their proposed features and values are presented in Table 3.

### Transport Plan

The transport plan is prepared according to the company's sales and procurement plan. It is important to consider the transport plan separately because it is often made from a shipper's perspective without considering the needs of transport service providers.

The time horizon reflects the optimal time horizon for transport planning from the shipper's perspective. For the railway, the longer the time horizon, the better.

The level of detail is an indicator of planning uncertainty. Transport plans that forecast demand on a monthly or weekly level are generally more uncertain than plans that specify daily demand. Highly aggregated forecasts are generally of little or no value for freight railways.

The revision cycle is another relevant feature. In the best case there would be no revisions (indicating stable transport forecasts), whereas daily revisions would be the worst case (indicating highly unpredictable transport demand). In the middle are weekly plan revisions and event-driven plan revisions (e.g., due to unexpected sales or production breakdowns). Event-driven plan revisions are almost as bad as daily plan revisions because they also make very-short-term changes to schedules.

### Transport Flexibility

The transport flexibility subcategory consists of physical and operational characteristics of the shipper that influence its ability to make accurate transport forecasts.

With the availability of transport alternatives at short notice, there is little incentive for shippers to create accurate long-term rail transport plans. Stable long-term rail transport plans can be developed if rail is used to meet basic transport demand, whereas other transport modes are used to meet peak demand. The reverse strategy would reduce forecast accuracy.

The latest possible time for placing a transport order reflects the fact that railways face strong competition from trucking, a mode that offers high scheduling flexibility, sometimes within hours. Therefore railways must also offer schedule flexibility to be competitive. The values for this feature range from 1 day or less to over 1 week. The shorter the allowable time horizon for ordering service, the less need for shippers to develop accurate forecasts.

The transport vehicle ownership feature considers the ability for shippers to adjust the number of freight wagons available for their use at any given time. The most accurate forecasts are made by shippers using their own exclusive wagon fleets (either owned or leased) because a limit is placed on the maximum possible number of shipped wagons or unit trains possible even if higher demand existed. Shippers operating their own wagon fleets also have a direct financial interest in maximizing fleet utilization, which is possible only with a stable and predictable transport schedule.

However, when shippers lease freight wagons on an ad hoc basis their schedule forecasts are less predictable. Ad hoc leasing is very common in single-wagonload services and can be arranged very quickly. In Switzerland, for example, shippers can order freight wagons 1 workday in advance (8). This ability enables shippers to adjust their shipping rapidly according to actual demand. Without incentives for early wagon reservations, shippers have little incentive to develop stable long-term transport plans.

The last feature is the availability of nontransport measures for addressing demand deviations. The more non-transport-related measures that are available, the more stable the transport plans. These measures differ significantly between and even within product sectors. An example of a non-transport-related measure is the ability to rent additional storage capacity on a short-term basis to

avoid the need for canceling a shipment if the company's own on-site storage facilities are full.

### *Shipper Forecast Information*

The third subcategory of transport planning is the quality of forecast information provided by the shipper to the freight railway. In this research forecasts are defined as nonbinding plans, whereas orders involve a financial penalty for changes. As shown in Table 3, there are five features of forecasts.

The most important feature is forecast accuracy. It is defined as the ratio of the number of changes of predicted shipments to the total number of predicted shipments. A change can be a cancellation, a change of destination, or a change in shipment date. Talks with production planning experts from a large Swiss freight railway revealed that forecasts having an accuracy of less than 80% are generally either not transmitted to freight railways or otherwise not considered by the railway in production planning.

The share of shipments forecast reflects the accuracy of forecasting the quantity of goods to be shipped. If the share of shipments forecast is >90%, this means that more than 90% of the shipped quantity of goods was actually forecast. Low percentages indicate that the shipper regularly underestimates the quantity of goods to be shipped. Generally overestimates are worse for the railway because that means resources are assigned but not needed.

Further, a high and regular forecast frequency is better for the railway because it means the railway has up-to-date information about expected future transport demand. Similarly, for the last two features in this category, a long forecast horizon and a high level of detail are best for predicting the quality of demand forecasts provided that accuracy remains acceptably high.

### *Shipper Order Information*

The final subcategory of information concerning transport planning is the quality of the order information. Although shippers do not want to change orders because they will incur a financial penalty, sometimes the change is necessary. A freight railway's acceptance of order changes and extra orders is a gesture of goodwill and an important negotiating point. As shown in Table 3 there are two main features of order information.

As with forecast information the most important factor for orders is their accuracy. The accuracy of shipment orders is defined as the ratio of the number of changes made to shipment orders + additional shipment orders divided by the total number of shipment orders placed.

The dominant type of order change feature reflects the fact that not all types of order changes have the same implications for railway scheduling. The most favorable changes are extra orders placed after the due date. The extra orders generally improve railway production plans because railways are free to reject these requests when they do not have adequate capacity.

Destination changes generally have minor implications for railway production planning as long as a major part of the initial itinerary remains unchanged. Changed shipping dates directly affect production planning, but their significance depends on the particular planning situation; therefore their impact is not generally assessable. Order cancellations are the worst possible situation for the railway because the revenue is lost but the costs for personnel and rolling stock (which cannot be redeployed in the short term) remain.

## CASE STUDY

The proposed framework was tested in several case studies with shippers in Switzerland. One of these case studies is exemplarily presented in this paper. The selected company ships refined mineral oil products (fuel and heating oil) in unit trains (primarily) from a refinery to large distribution depots in Switzerland. The shipping volume is approximately 20 loaded unit trains per week. This volume classifies the company as a large-scale shipper in Switzerland. The case study analysis is based on a face-to-face interview with the company's manager of rail transport planning as well as detailed forecast and order data from the company (covering all unit trains during 2009).

Table 4 summarizes the characteristics of forecasts and orders from the company. As shown in the table, the company forecasts are very helpful for the railway. More than 75% of total shipments are forecast on a detailed level, and these forecasts have a very high accuracy (more than 90%). The company regularly transmits forecasts to the freight railway at the end of each month for the entire upcoming month. During the month, there are few event-driven updates of this monthly forecast. This forecast quality is unsurpassed when compared with the forecasts of other companies shipping mineral oil in Switzerland. The accuracy of transmitted orders 1 week before shipping is close to 100%. The few subsequent changes of orders and forecasts primarily concern cancellations.

The research goal is to identify the features that help companies make accurate forecasts. Therefore the case study company's feature values in the proposed morphological framework should be clustered on the right side of the tables. Table 5 summarizes the feature values of the studied company.

The company has a reasonably complex transport network with two to five origins and more than 15 destinations. Furthermore, the percentage of transport volumes on the top 20% of routes is a little above 60%. This figure suggests a heterogenic distribution of transports through the network. The network extent is limited to Switzerland, which reduces complexity.

A problem for the company involves low sales planning accuracy reflected by the fact that the company relies on intuitive sales forecasts for heating oil produced on a monthly basis. The problem is that, by contrast to fuel demand, demand for heating oil is very difficult to predict. There are several reasons for this difficulty, including high customer price sensitivity, a time shift between sales and physical delivery of heating oil to end customers, and variations in weather. Therefore sales estimates of heating oil demand are made intuitively and require sound expert knowledge.

**TABLE 4** Forecast and Order Information Transmitted by Large Mineral Oil Company in Switzerland to Its Freight Railway

Subcategory	Feature	Case Study Company's Selected Value
Forecast information provided to freight railway	Forecast accuracy	>90%
	Share of shipments forecasted	75%–90%
	Forecast frequency	Monthly <sup>a</sup>
	Forecast time horizon	≥1 month
	Forecast level of detail	Day
Order information	Order accuracy	>95%
	Dominant type of order change	Cancel shipment

<sup>a</sup>Some event-driven updates of monthly forecast during months.

TABLE 5 Characteristics of Rail Transport Planning of Large Mineral Oil Company in Switzerland

Subcategory	Feature	Case Study Company's Selected Value
Transport network structure	Number of origins	2–5
	Number of destinations	>15
	Transport volume on top 20% of relations	60%–80%
	Network extent	Domestic or short distance
Transport responsibility	Number of railways involved	One
	Provider of shipping-paid deliveries	Yes
	User of shipping-paid deliveries	No
	Entity responsible for releasing products being shipped	= freight payer
Transport origin	Quantitative flexibility of production capacity at origin	Not flexible in terms of time
	Inventory philosophy at origin	Minimum inventory (just in time)
	Entity in charge of origin site	= entity placing shipment order
	Use of rail siding	Exclusive
Transport destination	Inventory philosophy at destination	Medium inventory
	Entity in charge of destination site	= entity placing shipment order
	Use of rail siding	Shared
Freight customer service characteristics	Shipment lot size	Unit trains
	Delivery date flexibility	±1 day
Sales and procurement plan	Time horizon	≥1 month
	Level of detail	Month
	Forecast technique	Intuitive <sup>a</sup>
		Mathematical <sup>b</sup>
Transport plan	Time horizon	≥1 month
	Level of detail	Day
	Revision cycle	Weekly
Transport flexibility	Availability of transport alternatives at short notice	No
	Latest possible time for placing transport order	≤1 week prior to shipping
	Transport vehicle ownership	Proprietary possession/long-term leasing
	Nontransport measures for addressing demand deviations	Many

<sup>a</sup>Heating oil.<sup>b</sup>Fuel.

However, there are several important reasons that this company is nevertheless able to forecast its rail transport demand accurately. First, the large majority of the company's rail transport departs from its own refinery. The refinery's flexibility of production capacity is limited, meaning that only minor modifications in production output are possible in a weekly time period. Furthermore the buffer function of production output depots at the refinery is low. This is supported by the fact that the inventory accounting at the company refinery is done at production input level only. Therefore, products must be shipped to distribution depots quickly, regardless of actual demand.

Second, the company's transport flexibility is limited. The company leases its rail tank wagons on long-term contracts. Therefore transport capacity remains constant within a period of at least several months and the company has an immediate financial interest to utilize its leased rail cars fully. This requires regular and stable shipment intervals, which are highly congruent with the interests of freight railways.

Fixed transport capacities also limit the maximum number of shipments at times of peak demand. This limit requires the company to find other solutions for coping with peak demand. The transport flexibility is also limited by a lack of transport alternatives because shipping large quantities of fuel and heating oil by truck is financially unattractive. Using other rail carriers is not an option in the case of this particular company.

Third, the company has numerous non-transport-related measures for coping with demand variations. This is explainable by statu-

tory stockpiling requirements for mineral oil companies and the fact that mineral oil companies are basically trading identical products. The quantity of statutory stockpiling for a mineral oil company is determined by law, but regulations allow bookkeeping transfers of statutory stocks between approved distribution depots in Switzerland.

This possibility allows the company's transport planning unit to address a shortage in one distribution depot by reducing that depot's compulsory stock and increasing the compulsory stock by the same amount in another depot. The fact that most distribution depots are operated as joint ventures and managed by the transport planning unit opens up a range of further possibilities. For instance, it enables transactions with competitors from the same distribution depot. These transactions can be transfers of stock or the short-term leasing of additional storage capacities to avoid an imminent overflow.

An additional advantage is that distribution depots generally store heating oil and fuel. Therefore, the analyzed company may operate unit trains as initially planned but with other goods (e.g., heating oil instead of fuel). It is also partly possible for the company to handle customer cancellations of shipping-paid unit train deliveries without changing transport plans. If these customers are using the same joint-venture distribution depots as the analyzed company, the already planned unit trains can be used to supply the company's own storage tanks. In both cases, no change in transport plans is necessary.

The above results are especially interesting when they are compared with results from other companies. This company serves as

a benchmark for a high forecast quality in the mineral oil sector of Switzerland. As the decisive underlying conditions have been identified, forecasts of a similar quality can be demanded from other companies with comparable characteristics.

## CONCLUSIONS

The morphological analysis scheme presented in this paper provides a structured framework for analyzing the quality of rail transport forecasting from freight shipping companies. The framework can help explain forecast and shipment order quality and also help analysts detect and understand inconsistencies between expected and actually provided forecast and order information.

The morphological scheme's main benefit is the possibility to compare rail transport planning from different freight shippers by using a uniform methodology. This methodology enables railways to determine benchmark values for achievable forecast periods and accuracy rates for freight shipping companies. Freight railways can use these benchmarks to decide which shippers to pursue for accurate forecasts, and for which the effort is not warranted. This approach is notably useful if forecast data are not available or only partly available from shippers.

Finally, the benchmark values can be used by other companies to help improve rail freight forecasting in their own companies.

Freight railways have an immediate interest in these improvements and should therefore push such activities actively.

## REFERENCES

1. Richey, T. General Morphological Analysis—A General Method for Non-Quantified Modelling. Presented at 16th EURO Conference on Operational Analysis, Brussels, Belgium, 1998.
2. Wissema, J. Morphological Analysis: Its Application to a Company TF Investigation. *Future*, Vol. 8, No. 2, 1976, pp. 146–153.
3. Schönsleben, P. *Integral Logistics Management*. St. Lucie Press, Boca Raton, Fla., 2004.
4. Meyr, H., and H. Stadtler. Types of Supply Chains. In *Supply Chain Management and Advanced Planning* (H. Stadtler and Ch. Kilger, eds.), Springer Verlag, Berlin, 2008.
5. Rohde, J., H. Meyr, and M. Wagner. Die Supply Chain Planning Matrix. *PPS Management*, Vol. 5, No. 1, 2000, pp. 10–15.
6. Fleischmann, B., H. Meyr, and M. Wagner. Advanced Planning. In *Supply Chain Management and Advanced Planning* (H. Stadtler and Ch. Kilger, eds.), Springer Verlag, Berlin, 2008.
7. Fries, N. *Market Potential and Value of Sustainable Freight Transport Chains*. Scientific Series of the Institute for Transport Planning and Systems, ETH Zürich, Zürich, Switzerland, 2010.
8. SBB Cargo. *Preise und Konditionen Güterwagen (Prices and terms of freight wagons)*. [http://www.sbbcargo.com/preise-konditionen-wagen\\_d.pdf](http://www.sbbcargo.com/preise-konditionen-wagen_d.pdf). Accessed July 22, 2011.

---

*The Freight Rail Transportation Committee peer-reviewed this paper.*

# Value for Railway Capacity

## Assessing Efficiency of Operators in Great Britain

Melody Khadem Sameni and John M. Preston

Growth in rail traffic has not been matched by increases in railway infrastructure. Given this capacity challenge and the current restrictions on public spending, the allocation and the utilization of existing railway capacity are more important than ever. Great Britain has had the greatest growth in rail passenger kilometers of European countries since 1996. However, costs are higher and efficiency is lower than European best practice. This paper provides an innovative methodology for assessing the efficiency of passenger operators in capacity utilization. Data envelopment analysis (DEA) is used to analyze the efficiency of operators in transforming inputs of allocated capacity of infrastructure and franchise payments into valuable passenger service outputs while avoiding delays. By addressing operational and economic aspects of capacity utilization simultaneously, the paper deviates from existing DEA work on the economic efficiency of railways by considering a new combination of input–output that also incorporates quality of service. The constant and variable returns to scale models are applied to the case study of franchised passenger operators in Great Britain. The follow-up Tobit regression model shows positive correlation between serving London and the efficiency scores. There is negative correlation between offering regional services (average length of journeys less than 40 mi) and the efficiency scores. The overall study and the results can provide helpful insights for railway authorities into the tactical and strategic planning of railways needed to increase efficiency.

Growth in rail passenger and freight traffic during the past decade has not been matched by increases in railway infrastructure capacity. Road congestion, higher fuel costs, privatization of railways, and concerns for sustainability and the environment are the main contributors to growth in rail demand. Because of infrastructure limitations, many railways worldwide are facing a capacity challenge to accommodate necessary train services on their infrastructure (1, 2). This situation is schematically shown in Figure 1. Increasing railway capacity by building new lines is extremely expensive and time-consuming. With the existing restrictions on public spending, it is only through increasing efficiency (more outputs by same level of or less inputs) that railways will remain sustainable in the medium term. For instance, the British railway industry is aiming for a 30% increase in its efficiency and annual savings of between £700 million (1£ = US\$1.55, 2011) and £1 billion by 2019 (3).

Infrastructure is the most valuable asset in the railway industry. When railway infrastructure is publicly owned and maintained (as in most European railways), allocating time slots for using the capacity

of infrastructure (train paths) to different train services is a critical task. In cases in which passenger operators are franchised by governments, it should be ensured that all these train paths and franchise payments are efficiently transformed into valuable socioeconomic outputs. However, until now there has been no holistic way of analyzing how well different passenger operators utilized the allocated capacity. This study for the first time in the literature focuses on externally obtained inputs, such as the allocated capacity of the infrastructure and the quality of service provided along with the quantity of outputs. The methodology is developed for the United Kingdom context; however, the approach can be applied in other countries for comparison.

### RESEARCH ON EFFICIENCY IN RAILWAYS

Efficiency is commonly assessed by the ratio of generated outputs to inputs (4). If it is considered in a wider context of value for money, it can be part of the chain of economy, efficiency, and effectiveness, or the three E's, as described by Booz & Company (5):

- Economy: how cheaply inputs are provided,
- Efficiency: quantity of output produced by using inputs, and
- Effectiveness: to what extent the money has delivered desired outputs.

The privatization of railways in Europe was followed by a wave of efficiency studies to research the effects of privatization on the efficiency of railways. The extensive survey conducted by Merkert et al. summarizes major works in this field (Table 1) (6).

Table 1 shows no studies that analyze efficiency in using allocated train paths to produce reliable and valuable services. In a broader sense, the focus of existing research has been on internally obtained inputs such as staff and rolling stock rather than on externally obtained inputs such as capacity of the infrastructure and franchise payments. The capacity of infrastructure is a limited resource, so how well train paths are allocated and used should be analyzed. Moreover, in the outputs, quality of service (e.g., delay minutes), which would be worthwhile to incorporate in the approach adopted in the current research, has never been considered. Figure 2 and Figure 3 compare the adopted approach of the current study with past approaches found in the literature.

In a typical vertically separated railway, objectives, interests, and concerns are segmented as well. The government, passenger operators, freight operators, and infrastructure authority have different responsibilities, objectives, and concerns. As a consequence of this segmentation, analyzing efficiency is highly dependent on who is chosen as the stakeholder. In this research government is considered to be the stakeholder responsible for the socioeconomic welfare of society. The efficiency of railway passenger operators is judged by

Transportation Research Group, Faculty of Engineering and the Environment, University of Southampton, Southampton SO17 1BJ, United Kingdom. Corresponding author: M. Khadem Sameni, m.sameni@soton.ac.uk.

*Transportation Research Record: Journal of the Transportation Research Board*, No. 2289, Transportation Research Board of the National Academies, Washington, D.C., 2012, pp. 134–144.  
DOI: 10.3141/2289-18



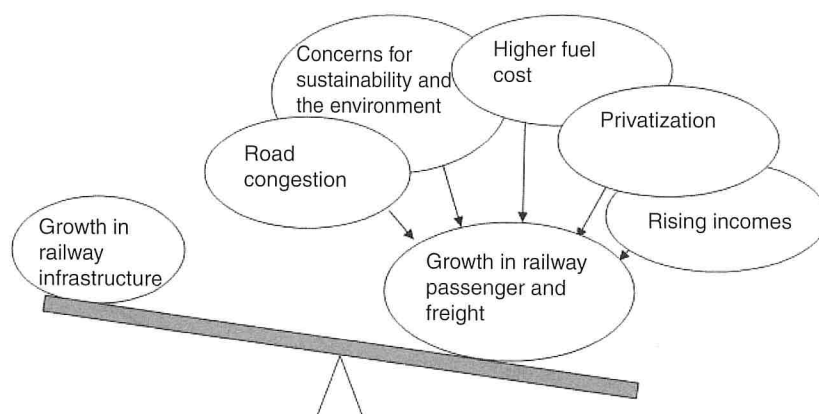


FIGURE 1 Scale of railway capacity challenge.

TABLE 1 Efficiency and Productivity Studies on Railways (24)

Study Reference	Method	Sample	Input	Output
(7)	Partial productivity measure (PPM)	14 European railways, 1970–1990	Staff train km; market share; receipts/total cost	
(8)	PPM	11 European railways, 1989–1994	Train km/track-km; train km/staff; market share; traffic units/train km; operating cost/train km; receipts/traffic units; revenue/costs	
(9)	Data envelopment analysis (DEA)	19 railways in Europe and Japan	Staff; energy consumption; rolling stock	Passenger km; freight-ton-km
(10)	Stochastic frontier analysis (SFA)	19 European railways, 1986–1988	Engines and railcars; staff, length of not electrified and electrified lines	Sum of passenger km and freight-ton-km
(11, 12)	DEA and corrected ordinary least squares (COLS)	17 European railways, 1988–1993	Staff; rolling stock; track length	Passenger km; freight-ton-km
(13)	SFA	16 European railways, 1970–1990	Operating cost; labor cost, energy, material/external	Passenger km; freight-ton-km
(14)	DEA	17 European railways, 1970–1995	Operating cost; track km	Passenger km; freight-ton-km
(15)	Total factor productivity (TFP)	10 European railways, 1969–1993	Staff; capital cost (interest and depreciation); energy cost	Sum of passenger km and freight-ton-km weighted with revenue share
(16)	PPM	14 railways in Europe and 5 American railways, 1977–1999	(Passenger km + freight-ton-km)/operating staff; traffic units/operating staff (1980–1999)	
(16)	SFA and TFP	14 railways in Europe and 5 American railways, 1977–1999	Staff; rolling stock (four categories)	Passenger km; freight-ton-km
(17)	PPM	15 railways worldwide, 2003–2004	(Passenger km + freight-ton-km)/total route length	
(18)	SFA	British train-operating companies (TOCs), 1996–2000	Staff; rolling stock; track length	Train km
(19)	DEA	54 railways in 27 countries, 2000–2004	Staff; rolling stock; track km; operating expenditure	Train km; passenger km; freight-ton-km
(20)	DEA	14 European railways, 1990–2001	Staff; track length; rolling stock	Passenger km; freight-ton-km
(21)	SFA	26 British TOCs, 1996–2006	Staff and rolling stock and other operating cost; wage prices, rolling stock characteristics; policy variables	Train km/route km, route km, vehicle km/train km
(22)	SFA	31 European railways, 1994–2005	Staff; rolling stock; network length	Passenger km; freight-ton-km
(23)	DEA	16 European rail systems, 1985–2004	Staff; rolling stock (passenger versus freight); network length	Passenger km; freight-ton-km
(6)	DEA	43 Swedish, German, and British train operating firms, 2006–2007	Material (annual amount spent on operation including depreciation and rolling stock lease costs but excluding all staff costs); total staff	Train km
			Material; managerial and administrative staff; remaining production staff	Train km; passenger km
			Material; managerial and administrative staff; remaining production staff	Train km; ton-km

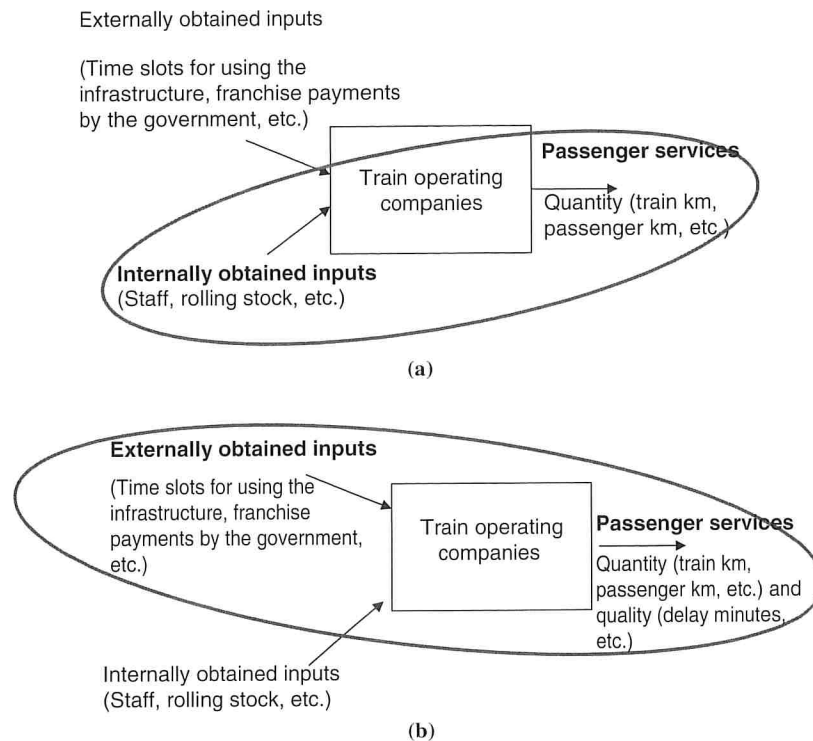


FIGURE 2 Comparing current research approach to efficiency analysis with past approaches: (a) past approaches to efficiency (TOC as stakeholder) and (b) approach of current research to efficiency (government as stakeholder).

the extent public resources are used as inputs to generate valuable passenger services for society. Aspects of quantity and quality of services are considered to assess the value of provided services. Figure 4 shows a schematic representation of inputs and outputs for analyzing operator efficiency as suggested by this paper. The efficiency of a train-operating company (TOC) is measured by its efficacy in converting the allocated capacity of infrastructure (time-tabled train kilometer) and financial support from the government (set largely by franchise agreements) into satisfactory passenger services. Discussion about the choice of inputs and outputs and the model will be provided in detail later in this paper.

## EFFICIENCY IN GREAT BRITAIN'S RAILWAY NETWORK

About 1.3 billion passenger journeys are made and 100 million metric tons of freight are transported on Britain's railway network annually and 1 million more trains run on Britain's railway network compared

with volumes in 2006. It is also forecast that passenger demand for rail will double and freight demand will increase by 140% during the next 30 years (25). Quality of services has considerably improved as well. The public performance measure (PPM) is the index that is usually used to reflect the quality of service which is "the percentage of passenger trains that arrive at their destination on time (not later than 5 min for local services and not later than 10 min for interurban trains). If a train is cancelled or is later than the threshold, it has not met the criteria" (25). The PPM for the year ending January 8, 2011, was 90.8% as compared with 78% 10 years ago (26). However, these achievements have incurred extensive costs. A recent study by Lovell et al. contends that "Britain's rail infrastructure manager faces an efficiency gap of 40 per cent against European best practice and that train operating costs have also risen substantially, both because of rising factor prices (wages and fuel) and because of deteriorating productivity" (27). Although quality and quantity of services have been considerably enhanced during the past decade, Figure 5 shows a breakdown of costs in Great Britain's railway network as well as the actual financial flows. The infrastructure provided by Network Rail accounts for

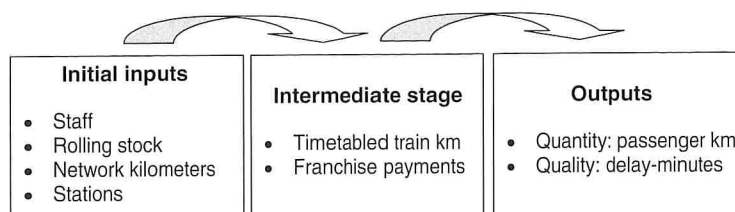


FIGURE 3 Transformation of inputs into outputs by TOCs and adopted approach of current research.

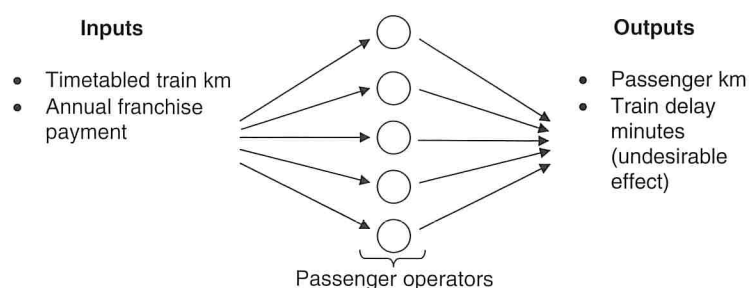


FIGURE 4 Inputs and outputs for analyzing operators' efficiency in capacity utilization.

the largest proportion of costs of the railway industry in Great Britain, making allocated capacity of infrastructure an invaluable resource. Improving the efficiency of utilizing the infrastructure, along with increasing the efficiency of train operations, is a robust means of decreasing costs to achieve the targeted annual cost saving of up to £1 billion as set by the Department for Transport and Office of Rail Regulation (3). The benchmarking study of four European railways by Civity Management Consultants suggests that Great Britain has the most competitive market structure and that market shares are distributed relatively widely among different operators (Figure 6), whereas in Sweden and the Netherlands state-owned companies still have market dominance (24). However, given the large costs in Great Britain, analyzing passenger train operator efficiency is highly important.

## DATA ENVELOPMENT ANALYSIS

Measuring productivity and efficiency in service industries (as opposed to manufacturing industries) is very challenging (30). The task is more complicated when a partnership of public and private resources is involved (as for many passenger railway services in Europe) as opposed to a purely private service industry (such as banking). Many elements are involved, for example, safety of operation, quality of service (delays, reliability, and stability), overall passenger satisfaction, and fare. Some of them are difficult to monetize. Moreover, they have different units of measurement and comparing them in a holistic manner is not easy. Data envelopment analysis (DEA), as a nonparametric tool based on linear programming, “provides a satisfactory measure of efficiency that takes into

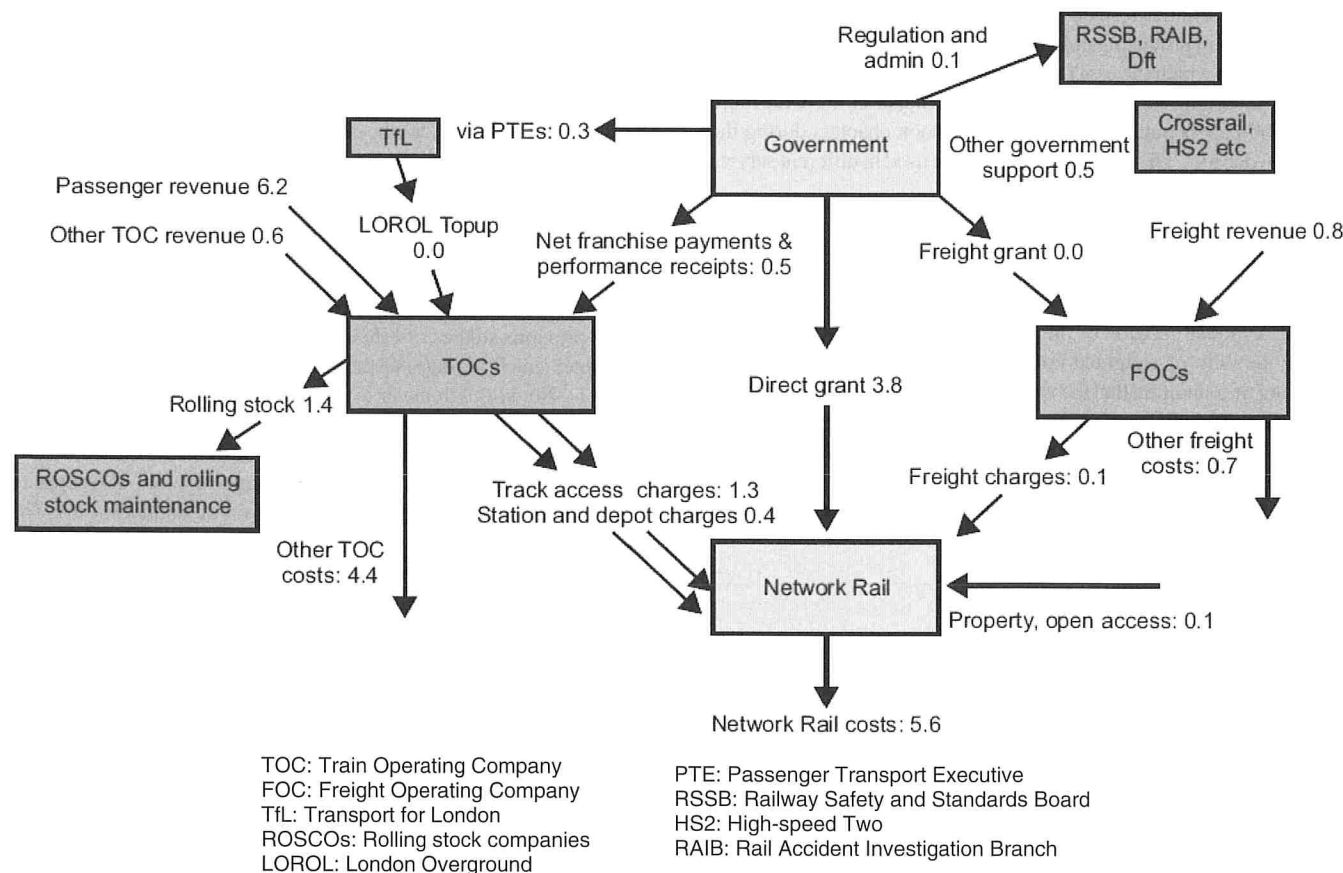


FIGURE 5 Cost breakdown and financial flows in GB rail 2009–2010 (£ billions) (3, 28, 29).

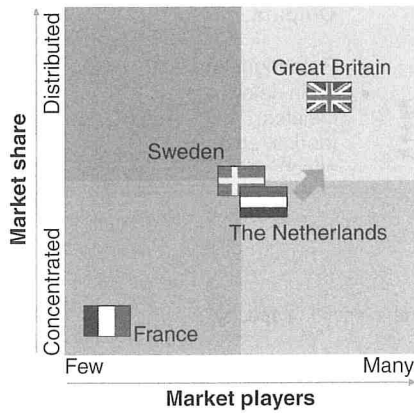


FIGURE 6 Market structure railways in four European countries (11).

account all inputs yet avoiding index number problems” (31). The co-winner of the Nobel prize for Economics in 1969, Ragnar Frisch, has described it as follows:

The index-number problem arises whenever we want a quantitative expression for a complex that is made up of individual measurements for which no common physical unit exists. The desire to unite such measurements and the fact that this cannot be done by using physical or technical principles of comparisons only, constitutes the essence of the index-number problem and all the difficulties centre here. (32)

DEA builds a frontier formed by the most efficient decision-making units (DMUs) in producing outputs from inputs. It is especially useful in instances in which the exact relationship between the outputs is not known. According to the survey by Emrouznejad et al., with more than 4,000 papers published in journals or book chapters during the past three decades, DEA has extensively been used in different service industries to analyze efficiency (33).

### The Model

The DEA model maximizes the efficiency of each DMU by maximizing the ratio of weighted outputs to weighted inputs subject to satisfying the condition that the weights are positive and that for every DMU, the efficiency score is less than or equal to unity. Considering  $n$  DMUs (TOCs),  $m$  inputs and  $s$  outputs,  $\epsilon$  as non-Archimedean infinitesimal,  $x_{ij}$  as the input  $i$  for DMU  $j$ ,  $y_{rj}$  as the output  $r$  for DMU  $j$ ,  $u$  and  $v$  as the weights for outputs and inputs, the formulation as suggested by Charnes et al. would be (34)

$$\max h_o = \frac{\sum_{r=1}^s u_r y_{ro}}{\sum_{i=1}^m v_i x_{io}}$$

$$\frac{\sum_{r=1}^s u_r y_{rj}}{\sum_{i=1}^m v_i x_{ij}} \leq 1 \quad j = 1, \dots, n$$

$$u_r, v_i \geq \epsilon \quad r = 1, \dots, s \quad i = 1, \dots, m$$

The above model is a fractional programming, and the linear version of the above formulation is

$$\min g_o = \sum_{i=1}^m \omega_i x_{io}$$

$$\sum_{i=1}^m \omega_i x_{ij} - \sum_{r=1}^s \mu_r y_{rj} \geq 0$$

$$\sum_{r=1}^s \mu_r y_{ro} = 1$$

$$\mu_r, \omega_i \geq 0$$

where

$h_o$  = efficiency of unit under assessment,  
 $u_r$  = weight given to output  $r$ ,  
 $y_{ro}$  = amount of output  $r$  for unit under assessment,  
 $v_i$  = weight given to input  $i$ ,  
 $x_{io}$  = amount of input  $i$  for unit under assessment,  
 $g_o$  = efficiency of the unit under assessment,  
 $\omega_i$  = weight given to input  $i$  in the linear model, and  
 $\mu_r$  = weight given to output  $r$  in the linear model.

### Choosing Inputs: Externally Obtained Resources

In DEA, “Inputs are defined as resources utilized by the DMUs or conditions affecting the performance of DMUs” (35). Timetabled train kilometer is used as a proxy to reflect infrastructure utilization by a train operator. The more trains that are run on the infrastructure and the longer distances they travel, the more capacity is used. This bigger input provides the chance to generate more valuable outcomes. It is worth emphasizing that the choice of inputs and outputs depends on the process being analyzed. Therefore, as analyzing efficiency of capacity utilization is the main aim of this study, unlike the studies mentioned in Table 1, train kilometer is an input in this study. The efficiency of the operators is analyzed in regard to transforming this allocated capacity of infrastructure into passenger services. Some previous studies of efficiency in railways have used network kilometer (track or route) as their input for DEA models (as seen in Table 1). Network kilometer is not an exact input to reflect capacity utilization, which is the main goal of this study. It depends on how many trains run on these routes. If no train runs on the infrastructure, capacity utilization would be zero according to the UIC 406 capacity leaflet developed by the International Union of Railways (36). The higher the number of trains that run on the infrastructure in the time unit, the higher the capacity utilization index would be.

Franchise payments by government are an external input that can be used for analyzing operator efficiency in capacity utilization and converting them into valuable train services. It is a public resource that should be used as efficiently as possible.

### Choosing Outputs: Public Value of Services Provided

Outputs are the benefits generated as a result of the operations of the DMUs (35). The value generated by a passenger train operator can be measured in different ways. The first option that comes

to mind is to consider the revenue that is generated through ticket sales. However, this option cannot be a good index for operational efficiency. Trains running with a low load factor but high fares might be economically efficient but they are not operationally efficient.

Passengers transported (number of passenger journeys) is not by itself informative: one passenger might use the train for a very short distance; one might take the train for a very long distance. Therefore the best measure for considering the value generated by a train operator through using the infrastructure is to consider passenger kilometer. Passenger kilometer is also a good indicator for considering the environmental effects of railways as the carbon dioxide emissions saved by choosing railways as the mode of transportation is likely to be proportional to passenger kilometer (along with other factors such as train loadings, mode switching, and traction energy source).

Considering timetabled train kilometer as input and passenger kilometer as one of the outputs also covers both aspects of "macro" and "micro capacity utilization" as well as the concept of "lean capacity utilization" as suggested by Khadem Sameni et al. (37).

There is a trade-off between railway capacity utilization and quality of service. Therefore quality of service should be considered when a proper capacity analysis is provided. For each train operator company, a wide range of data on quality of service is available:

- Number of complaints received per 100,000 passenger journeys,
- National Passenger Survey results (detailed survey on quality of services on board trains and at stations carried out twice per year by Passenger Focus),
- PPM, and
- Delay minutes.

The number of complaints is not a good indicator for quality of service to be included in the DEA model. Complaints can be subjective and originate mostly from train performance. As indicated by the Office of Rail Regulation, in the financial year 2009–2010, 36% of total complaints were on train service performance; 21% on fares, retailing, and refunds; and 12% on quality of the train (38). Therefore a train performance indicator would be a better estimate of quality of service provided by the operator. Quality of services on board and at stations matters, but the first priority of passengers is getting to their destination on time. The PPM is a relative index, so instead an absolute measure of delay minutes was chosen to indicate the quality of service that is important for both passenger and network owner. This choice is also in line with the work of Tongzon, who used delay time (the difference between total berth time plus time waiting to berth and the time between the start and finish of ship working) for analyzing the maritime industry through DEA (39). Data on train delay minutes for different operators are provided in the *Annual Return* published by Network Rail (40).

Train delays are not a positive outcome. Therefore, in DEA terminology they are called "undesirable effects" and cannot be used directly in the model as outputs. Methods to handle them have been surveyed by Seiford and Zhu (41). The most popular methods are transferring undesirable effects to the input side (as DEA tries to minimize use of inputs) or using the inverse of "undesirable effects" as outputs (as DEA tries to maximize outputs). The second approach is used in this paper.

## Analysis of Results

DEA models can have two general orientations: input oriented or output oriented. The input-oriented model tries to minimize inputs

while at least the given level of outputs are produced, whereas the output-oriented model tries to maximize outputs while no more than the observed level of inputs are used (4). Because the main aim of the study is to increase the efficiency of railways by cutting costs, the input-oriented model was chosen. The models for constant return to scale (CRS) and variable return to scale (VRS) were solved by using PIM DEA-V3.0 software (42). DEA efficiency scores are presented in Table 2.

TOCs that have the highest average train utilization (Figure 7) tend to obtain higher efficiency scores. For example, East Coast and Virgin Trains, which carry the highest number of passengers per train, have also received the highest efficiency scores by the DEA model. However when delay minutes and franchise payments are considered, the ranking is not exactly the same because a TOC might not have performed well enough to provide punctual services or be cost-efficient. For instance, First Great Western has the third rank according to the average train utilization but ranks fifth when quality of service provided and franchise payments received are considered by the DEA model. To gain 100% relative efficiency, target values as suggested by DEA are used, as shown in Table 3. They are calculated by the PIM DEA-V3.0 software on the basis of the distance of the production possibility set from the efficient frontier as a result of benchmarking the relationships between inputs and outputs (42).

Operators that are less efficient are consuming more than necessary infrastructure capacity (reflected by timetabled train kilometer) to generate passenger kilometers or are less efficient in producing on-time services or receive greater than necessary franchise payments. Reducing nonefficient timetabled train kilometers (that are not transformed into passenger kilometers efficiently) would increase train load factors and the efficiency of capacity utilization. Such a reduction would also be likely to reduce delays. Reducing ineffective subsidies would increase operational efficiency.

Table 3 can provide insights to railway practitioners about how train operators can improve their operational efficiency. The operators that are not operationally efficient, as indicated by Smith, might (44)

- Have a low overall load factor,
- Haul a lot of empty seats off peak, and
- Haul empty seats long distances to satisfy short-distance demand.











Some of the ways that operators can increase their efficiency are by

- Reducing the frequency of their trains to increase their load factor (passenger per train) [for example, the East Coast operator that had the highest relative efficiency has the highest ratio for passenger journeys per trains planned (413.0) and the highest ratio for passenger kilometers per timetabled train kilometer (228.6)]. These ratios were 82.9 and 43.5 for Arriva Trains Wales and 99.9 and 43.4 for Northern.
- Using marketing techniques to attract more passengers to their current services and increase load factor.
- Trying to increase the reliability of their train services by reducing train delays.

The results can also provide helpful insights for railway authorities to plan better for infrastructure and franchise payments. For instance, the model results indicate a relatively low level of operational efficiency for the Cross Country services and the need for cuts in franchise payments and allocated timetable kilometers. By taking a closer look at the geographical area of operation for Cross Country trains, the overlap with four other TOCs that are operationally









TABLE 2 Efficiency Scores of Train Operating Companies in 2009

Name of Operator	Geographic Area of Operation (43)	CRS Model		VRS Model	
		Efficiency Score	Rank	Efficiency Score	Rank
Arriva Trains Wales		0.19	15	0.50	11
Chiltern Railways		0.53	6	1.00	1
Cross Country		0.40	7	0.48	14
East Coast		1.00	1	1.00	1
East Midlands Trains		0.39	11	0.59	8
First Capital Connect		0.56	4	0.71	6
First Great Western		0.54	5	0.97	5
First Scot Rail		0.27	14	0.35	15
London Midland		0.30	13	0.49	12
National Express East Anglia		0.50	7	0.54	9

(continued)

TABLE 2 (continued) Efficiency Scores of Train Operating Companies in 2009

Name of Operator	Geographic Area of Operation (43)	CRS Model		VRS Model	
		Efficiency Score	Rank	Efficiency Score	Rank
Northern		0.19	15	0.34	16
South Eastern		0.50	8	0.53	10
Southern		0.47	9	0.49	13
South West Trains		0.59	3	1.00	1
Trans Pennine Express		0.35	12	0.64	7
Virgin Trains		0.65	2	1.00	1

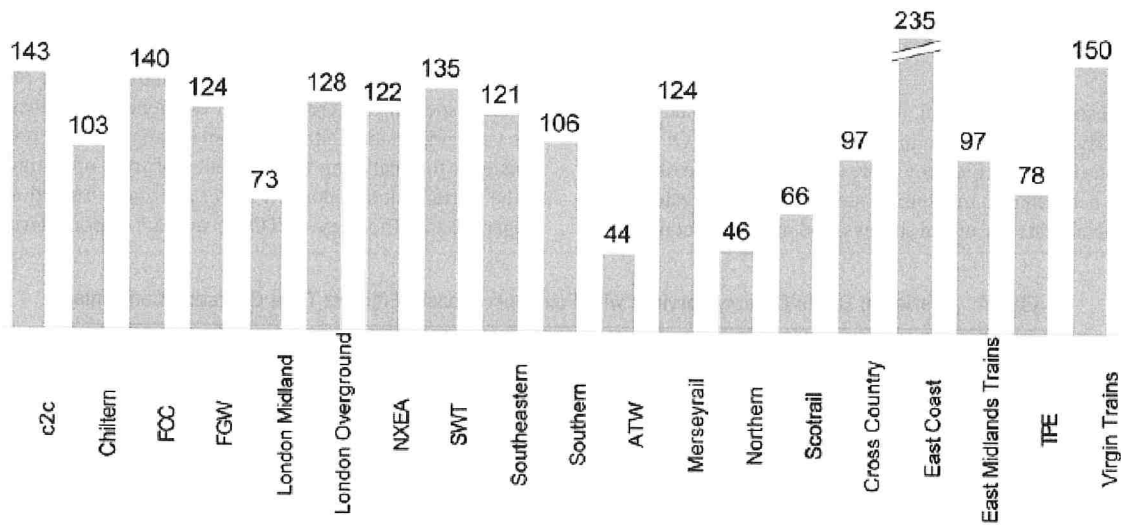


FIGURE 7 Average number of passengers per timetabled train (24).

TABLE 3 Target Values as Suggested by DEA

Name	Gain in Delay Minutes, 2009–2010 (%)	Gain in Passenger Kilometers (millions), 2009–2010 (%)	Gain in Timetabled Train Kilometers (millions), 2009–2010 (%)	Gain in Subsidy 2009–2010 (%)
Arriva Trains Wales	–61.38	71.86	–49.61	–49.61
Chiltern Railways	0	0	0	0
Cross Country	–69.99	0	–52	–71.51
East Coast	0	0	0	0
East Midlands Trains	–59.45	4.99	–40.95	–40.95
First Capital Connect	–40.92	15.6	–28.79	–28.79
First Great Western	–13.87	0	–17.63	–2.98
First Scot Rail	–78.52	0	–65.2	–73
London Midland	–75.95	7.33	–50.58	–50.58
National Express East Anglia	–72.77	0	–45.8	–49.48
Northern	–83.23	49.54	–66.03	–66.03
South Eastern	–70.68	0	–46.93	–81.47
Southern	–72.65	0	–50.94	–83.76
South West Trains	0	0	0	0
Trans Pennine Express	–63.58	3.12	–35.92	–35.92
Virgin Trains	–7.33	0	0	0

efficient can be seen (Table 4). Therefore, it is suggested that this is not an operationally efficient configuration. Alternative configurations might involve Cross Country services being divided between these other four TOCs or being operated by one of them (as was the case when they were operated by Virgin between 1997 and 2007) or by long-distance TOCs being merged. These initiatives could improve the efficiency of railways as targeted by the Department for Transport and Office of Rail Regulation in the value for money study (3).

Another point worth mentioning is the importance and role of the timetabling process. Heterogeneity of train speeds and mix of trains affect the performance of TOCs that share a specific route, but these are decisions that are largely external to the TOCs. The minimum number of train services for each franchise contract is specified by the Department for Transport, and timetables are devised by Network Rail in consultation with TOCs.

### Tobit Regression

In the second stage of the model, a Tobit regression is used to analyze the underlying factors affecting operator efficiency. Of interest are correlations between efficiency scores and average age of rolling stock, public performance measure, route kilometers operated, passenger satisfaction rates in annual surveys, and number of complaints

received. Tobit regression was done for the VRS model with SPSS V.19, by adding R and Python plug-ins and the SPSSINC\_TOBIT\_REGR application (45). The results for the Gaussian (normal) assumption are presented in Table 5.

Table 5 shows that the efficiency score is positively correlated with serving London ( $P$  value < .003). Offering regular services to London was chosen as the criterion; therefore Scot Rail, which offers a sleeper service to London, received zero for this variable. The efficiency scores are negatively correlated with the average length of journeys for regional services ( $P$  value < .011). Services whose average journey length was less than 40 mi according to the National Rail Trends statistics were considered to be regional (38). The average age of rolling stock and the number of staff employed were found to be insignificant factors.

### CONCLUSIONS

In the face of the railway capacity challenge and restrictions on public spending, increasing the efficiency of railway operations is very important. Data envelopment analysis can provide helpful insights for analyzing the efficiency of train operating companies. The current study adopts a novel approach toward analyzing the operational efficiency of TOCs; instead of considering internally

TABLE 4 Overlap of Cross Country Services with Four Operationally Efficient Train Operating Companies






Name of Operator	Cross Country	East Coast	Virgin Trains	Southwest Trains	First Great Western
Geographical area of operation					
Efficiency score	0.48	1	1	1	0.97

TABLE 5 Tobit Regression Results

Variable	Coefficient	Standard Error	z-Value	Sig.
Intercept	.664	.159	4.181	.000
Serving London	.332	.110	3.007	.003
Regional (average length of journeys less than 40 mi)	-.294	.115	-2.558	.011
Average age of rolling stock	-.003	.009	-.314	.754
Number of employees, 2009–2010	.000	.000	.076	.940
Log (scale)	-1.648	.217	-7.578	.000

NOTE: Scale = 0.1925; residual degrees of freedom (d.f.) = 10; log likelihood = -7.098, d.f. = 6; Wald statistic = 17.566, d.f. = 4; sig. = significance.

obtained resources, the current study considers externally obtained public resources, such as franchise payments and allocated capacity of the infrastructure.

Operational efficiency of train operating companies in transforming franchise payments and timetabled train kilometers to passenger kilometers while avoiding delays (an undesirable effect) is quantified by using DEA, and the model can suggest optimum values for inputs or outputs. Also some insights can be provided for strategic and tactical planning in cases in which the routes operated by an operator are not efficient and reconfigurations can be suggested. For example, routes operated by a less operationally efficient TOC that overlap with services of other more operationally efficient train operators could be reallocated between them to increase utilization of the value for railway capacity. A low efficiency score for a TOC does not necessarily imply that the TOC is incompetent. Franchise specification, service types (intercity versus regional), population density, and travel demand (e.g., road congestion) affect the efficiency scores, and these factors vary across different TOCs.

Increasing the efficiency of TOCs is just one of the ways to improve railway throughput. Improving other aspects of track capacity utilization, including better maintenance of the infrastructure to improve track condition, investing in new lines, enhancing junctions (e.g., flyovers) and stations, and modernizing signaling and train control, would enhance railway throughput.

The follow-up Tobit regression shows that high efficiency scores are strongly associated with serving London, whereas low scores are associated with short-haul regional services. The number of employees and average age of rolling stock do not have a statistically significant effect on efficiency scores.

## ACKNOWLEDGMENTS

The authors thank Alex Landex of the Technical University of Denmark for his comments on the first version of this paper and Ali Emrouznejad of Aston University for answering many data envelopment questions.

## REFERENCES

1. *Ten-year European Rail Growth Trends*. Association of Train Operating Companies, London, 2007.
2. Cambridge Systematics. *National Rail Freight Infrastructure Capacity and Investment Study*. Association of American Railroads, Cambridge, Mass., 2007.
3. Department for Transport and Office of Rail Regulation. *Realising the Potential of GB Rail—Final Independent Report of the Rail Value for Money Study*. London, 2011. <http://www.dft.gov.uk>.
4. Cooper, W. W., L. M. Seiford, and K. Tone. *Data Envelopment Analysis: A Comprehensive Text with Models, Applications, References and DEA-Solver Software*. Springer, New York, 2006.
5. Booz & Company. *Rail Value for Money Study: Research on Value for Money Assessment*. Commissioned by Department for Transport and Office of Rail Regulation. London, 2011. [www.dft.gov.uk](http://www.dft.gov.uk).
6. Merkert, R., A. S. J. Smith, and C. A. Nash. Benchmarking of Train Operating Firms—A Transaction Cost Efficiency Analysis. *Transportation Planning and Technology*, Vol. 33, No. 1, 2010, pp. 35–53.
7. Nash, C., and J. Preston. Railway Performance—How Does Britain Compare? *Public Money and Management*, Vol. 14, No. 4, 1994, pp. 47–53.
8. Nash, C. A., and J. D. Shires. Benchmarking of European Railways: An Assessment of Current Data and Recommended Indicators. In *Methodologies, Applications and Data Needs*, Transport benchmarking. European Conference of Ministers of Transport and European Commission, OECD, Paris, 2000, pp. 119–136.
9. Oum, T. H., and C. Yu. Economic Efficiency of Railways and Implications for Public Policy: A Comparative Study of the OECD Countries' Railways. *Journal of Transport Economics and Policy*, Vol. 28, No. 2, 1994, pp. 121–138.
10. Gathon, H. J., and P. Pestieau. Decomposing Efficiency into Its Managerial and Its Regulatory Components: The Case of European Railways. *European Journal of Operational Research*, Vol. 80, No. 3, 1995, pp. 500–507.
11. Coelli, T., and S. Perelman. A Comparison of Parametric and Non-Parametric Distance Functions: With Application to European Railways. *European Journal of Operational Research*, Vol. 117, No. 2, 1999, pp. 326–339.
12. Coelli, T., and S. Perelman. Technical Efficiency of European Railways: A Distance Function Approach. *Applied Economics*, Vol. 32, No. 15, 2000, pp. 1967–1976.
13. Cantos, P., and J. Maudos. Regulation and Efficiency: The Case of European Railways. *Transportation Research Part A: Policy and Practice*, Vol. 35, No. 5, 2001, pp. 459–472.
14. Cantos, P., J. M. Pastor, and L. Serrano. *Cost and Revenue Inefficiencies in the European Railways*. Instituto Valenciano de Investigaciones Económicas, 2002.
15. Loizides, J., and E. G. Tsionas. Dynamic Distributions of Productivity Growth in European Railways. *Journal of Transport Economics and Policy*, Vol. 38, No. 1, 2004, pp. 45–75.
16. Rivera-Trujillo, C. *Measuring the Productivity and Efficiency of Railways (an International Comparison)*. University of Leeds, Leeds, United Kingdom, 2004.
17. Hatano, L. The International Benchmarking Project. *Railway Strategies*, Jan.-Feb. 2005, pp. 70–72.
18. Cowie, J. Technical Efficiency Versus Technical Change—The British Passenger Train Operators. In *International Conference on Competition Ownership in Land Passenger Transport*, Elsevier, Amsterdam, Netherlands, 2005.
19. Growitsch, C., and H. Wetzel. Testing for Economies of Scope in European Railways: An Efficiency Analysis. *Journal of Transport Economics and Policy*, Vol. 43, No. 1, 2009, pp. 1–24.
20. Driessen, G., M. Lijesen, and M. Mulder. *The Impact of Competition on Productive Efficiency in European Railways*. CPB Netherlands Bureau for Economic Policy Analysis, 2006.
21. Smith, A., and P. Wheat. A Quantitative Study of Train Operating Companies Cost and Efficiency Trends 1996 to 2006: Lessons for Future Franchising Policy. European Transport Conference, 2007.
22. Wetzel, H. European Railway Deregulation: The Influence of Regulatory and Environmental Conditions on Efficiency. Institute of Economics, University of Lüneburg, Lüneburg, Germany, 2008.
23. Cantos, P., J. M. Pastor, and L. Serrano. Vertical and Horizontal Separation in the European Railway Sector and Its Effects on Productivity. *Journal of Transport Economics and Policy*, Vol. 44, No. 2, 2010, pp. 139–160.
24. Civity Management Consultants. *International Whole Industry Including Train Operating Cost Benchmarking*. Commissioned and published online by Department for Transport and Office of Rail Regulation, Hamburg, Germany, 2011.
25. *Britain Relies on Rail*. Network Rail, London, 2011.
26. Network Rail. *How Do We Measure Up?* 2011. <http://www.networkrail.co.uk/aspx/699.aspx>.

27. Lovell, K., C. Bouch, A. Smith, C. Nash, C. Roberts, P. Wheat, C. Griffiths, and R. Smith. Introducing New Technology to the Railway Industry: System-Wide Incentives and Impacts. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, Vol. 225, No. 2, 2100, pp. 192–201.
28. Atkins. *Rail Value for Money Study: Asset Management and Supply Chain Management Assessment of GB Rail*. Commissioned by Department for Transport and Office of Rail Regulation. London, 2011. [www.dft.gov.uk](http://www.dft.gov.uk).
29. Arup. *Rail Value for Money Study: Rolling Stock Whole Life Costs*. Commissioned by Department for Transport and Office of Rail Regulation. London, 2011. [www.dft.gov.uk](http://www.dft.gov.uk).
30. Bryson, J. R., and P. W. Daniels. *The Handbook of Service Industries*. Edward Elgar Publishing, Cheltenham, United Kingdom, 2007.
31. Farrell, M. J. The Measurement of Productive Efficiency. *Journal of the Royal Statistical Society, Series A (General)*, Vol. 120, No. 3, 1957, pp. 253–290.
32. Frisch, R. Annual Survey of General Economic Theory: The Problem of Index Numbers. *Econometrica: Journal of the Econometric Society*, 1936, pp. 1–38.
33. Emrouznejad, A., B. R. Parker, and G. Tavares. Evaluation of Research in Efficiency and Productivity: A Survey and Analysis of the First 30 Years of Scholarly Literature in DEA. *Socio-Economic Planning Sciences*, Vol. 42, No. 3, 2008, pp. 151–157.
34. Charnes, A., W. W. Cooper, and E. Rhodes. Measuring the Efficiency of Decision Making Units. *European Journal of Operational Research*, Vol. 2, No. 6, 1978, pp. 429–444.
35. Ramanathan, R. *An Introduction to Data Envelopment Analysis: A Tool for Performance Measurement*. Sage Publications Pvt. Ltd., London, 2003.
36. *Capacity (UIC Code 406)*. International Union of Railways (UIC), Paris, 2004.
37. Khadem Sameni, M., A. Landex, and J. Preston. Developing the UIC 406 Method for Capacity Analysis. Presented at 4th International Seminar on Railway Operations Modelling and Analysis, Rome, 2011.
38. Office of Rail Regulation. *National Rail Trends 2009–2010*. 2010. <http://www.rail-reg.gov.uk>. Accessed Oct. 5, 2011.
39. Tongzon, J. Efficiency Measurement of Selected Australian and Other International Ports Using Data Envelopment Analysis. *Transportation Research Part A: Policy and Practice*, Vol. 35, No. 2, 2001, pp. 107–122.
40. Network Rail. *Annual Return 2010 Report*. 2010. <http://www.networkrail.co.uk>. Accessed April 4, 2011.
41. Seiford, L. M., and J. Zhu. Modeling Undesirable Factors in Efficiency Evaluation. *European Journal of Operational Research*, Vol. 142, No. 1, 2002, pp. 16–20.
42. Emrouznejad, A., and E. Thanassoulis. *Performance Improvement Management Software*. 2011. <http://www.deasoftware.co.uk>.
43. Network Rail. *Train Operating Companies*. 2011. [http://www.nationalrail.co.uk/tocs\\_maps/tocs](http://www.nationalrail.co.uk/tocs_maps/tocs).
44. Smith, D. Timetabling Is the Culprit in ‘Empty Seats’ Problem. In *Rail Professional*, Rail Professional Ltd, Cambridge, United Kingdom, 2009.
45. IBM. IBM developers website. 2011. <http://www.ibm.com/developerworks>.

---

*The Freight Rail Transportation Committee peer-reviewed this paper.*



# Development and Assessment of Taxonomy for Performance-Shaping Factors for Railway Operations

Miltos Kyriakidis, Arnab Majumdar, Gudela Grote, and Washington Y. Ochieng

Human performance is a significant contributor to railway incidents and accidents. The literature shows that train drivers, signallers, and controllers most affect network safety. Several studies have been conducted in the field of human factors and human performance in the railway domain to investigate operators' influence on the railway system. However, most studies are based on previous studies from other domains, which are not well suited and can be difficult to apply reliably to railway-specific operations. In light of the current limitations, this paper proposes a new approach referred to as the human performance railway operational index (HuPeROI). HuPeROI aims not only to estimate the human error probability for railway operations but also to propose mitigation strategies to minimize phenomena such as operators' degraded performance. HuPeROI is based on a performance-shaping factors taxonomy designed for the rail industry, referred to as the railway performance-shaping factor (R-PSF) taxonomy. The R-PSF taxonomy was developed on the basis of an extensive literature review of the field of human factors and subsequently validated against the findings derived from the analysis of 179 railway accident and incident reports, as well as targeted interviews with subject matter experts. This paper presents the taxonomy and the underlying theory, as well as the results of the validation process based on the analysis of 179 reports and the assessment of R-PSFs for four different scenarios based on a subject matter expert elicitation process. The R-PSF taxonomy and HuPeROI should enable researchers to address and deal with the effect of degraded performance of railway operators.

The railway system is a major component of the economies of most countries, by daily transporting millions of passengers as well as millions of dollars worth of goods from origin to destination (1). The relevant operational, regulatory, and governmental bodies of every country with a rail network aim for a highly reliable, excellent quality, and safe railway system (2). Although the definition of a railway system can be broad, in this paper the definition includes only infrastructure, rolling stock, and rail employees.

The safety of railway operations in this system depends on several factors, including rail traffic rules, infrastructure reliability, organiza-

tional safety culture, and human factors (3). In recent years, interest in the area of human factors in railway operations has increased significantly (4). Again, although there are numerous definitions of human factors, the United Kingdom's Health and Safety Executive definition is adopted in this paper: "environmental, organisational, job factors, and human and individual characteristics, which influence behaviour at work in a way, which can affect health and safety" (5).

It is well recognized that a large number of railway accidents occur as a result of degraded human performance (1). Human performance, which can be either positive or negative, can be described as the human capabilities and limitations that have an effect on the safety and efficiency of operations (6). However, researchers are usually interested in negative human performance. A recent study shows that at least 75% of fatal railway accidents in Europe between 1990 and 2009 were caused by human error, for example, exceeding speed, signal passed at danger, or signaling or dispatching error (7).

The literature shows that it is the train drivers, signallers, and controllers (referred to as operators) who most affect network safety (1). Several studies have been conducted in the field of human factors and human performance in the railway domain (2). However, most of these studies are based on previous studies in the field of human reliability analysis from other domains, which are not well suited to the rail industry and can be difficult to apply reliably to railway-specific operations (8).

In light of the current limitations, this paper presents a new approach referred to as the human performance railway operational index (HuPeROI). It aims not only to estimate the human error probability for railway operations but also to propose mitigation strategies to minimize phenomena such as operators' degraded performance. (HuPeROI considers any "train movement from one point to another or during a shunting operation" to be a railway operation.) Maintenance or design personnel are not included in this index, and furthermore accidents or incidents due to passengers, trespassers, or third parties, for example, level crossing accidents, are also ignored.

HuPeROI is based on a performance-shaping factor (PSF) taxonomy designed for the rail industry. This taxonomy, referred to as railway performance-shaping factors (R-PSFs), was initially developed on the basis of an extensive literature review in the field of human factors and subsequently validated against the findings derived from the analysis of railway accident and incident reports. In addition, an experts' elicitation process was performed to assess the R-PSFs for four different railway operational scenarios. An overview of the HuPeROI development as well as that of the R-PSF taxonomy is given in Figure 1.

This paper introduces the R-PSF taxonomy and presents the results of the reports analysis, together with the findings of the

---

M. Kyriakidis, A. Majumdar, and W. Y. Ochieng, Centre for Transport Studies, Department of Civil and Environmental Engineering, Imperial College London, London SW7 2AZ, United Kingdom. G. Grote, Organisation, Work, and Technology Group, Department of Management, Technology, and Economics, Eidgenössische Technische Hochschule Zurich, Kreuzplatz 5, Zurich 8032, Switzerland. Corresponding author: M. Kyriakidis, m.kyriakidis@imperial.ac.uk.

*Transportation Research Record: Journal of the Transportation Research Board*, No. 2289, Transportation Research Board of the National Academies, Washington, D.C., 2012, pp. 145–153.  
DOI: 10.3141/2289-19

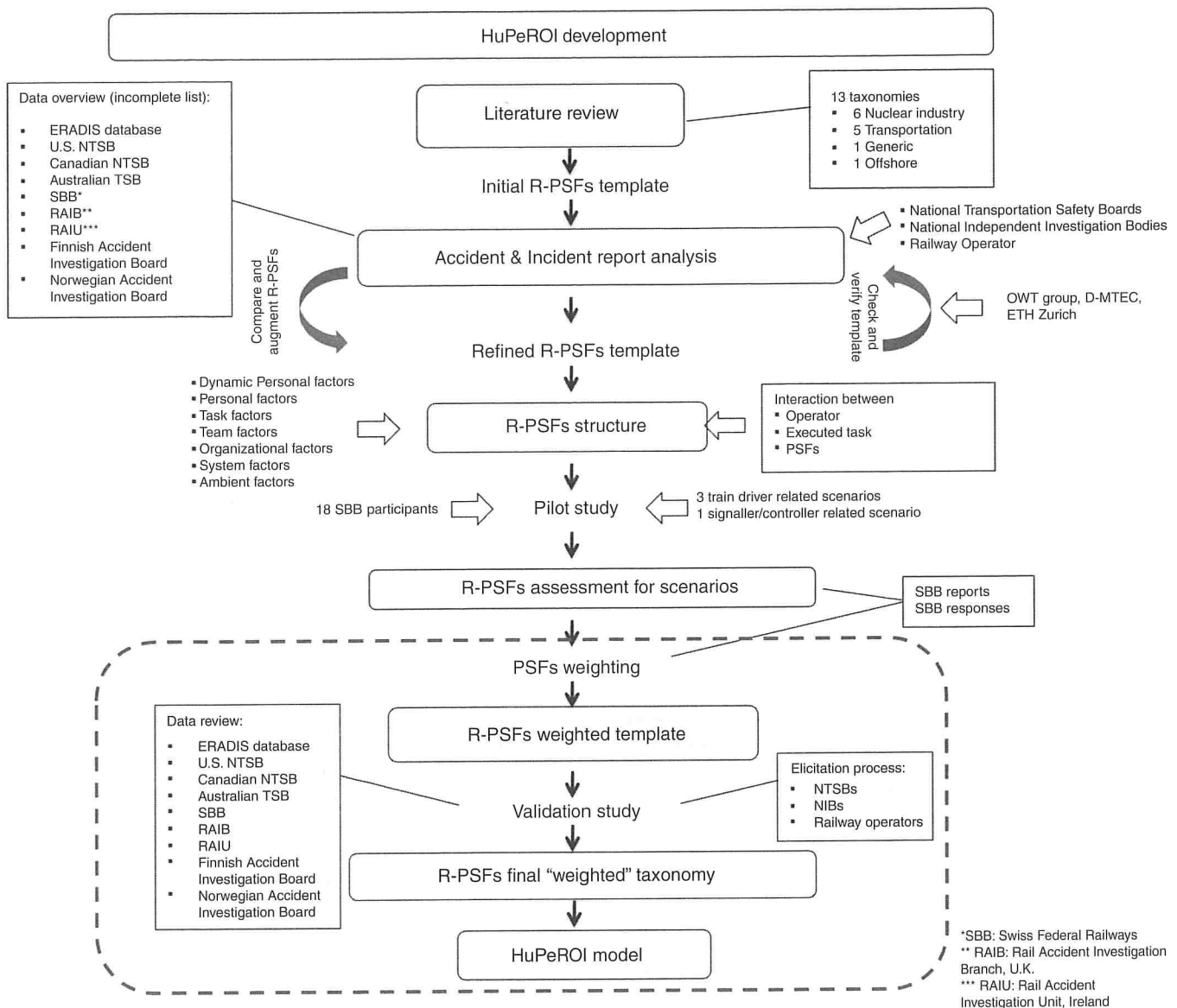


FIGURE 1 Overview of HuPeROI and R-PSF taxonomy development (ERADIS = European Railway Agency Database of Interoperability and Safety; NIBs = national independent investigation bodies; NTSB = National Transportation Safety Board; TSB = Transport Safety Bureau; OWT = Organisation, Work, and Technology; D-MTEC = Department of Management, Technology, and Economics; ETH = Eidgenössische Technische Hochschule).

elicitation process. The next section of the paper describes the current state of PSFs in several domains, followed by a section that presents the R-PSF taxonomy and the underlying theory. Results of the analysis of accidents and incidents data are presented next, in addition to the pilot study of the assessment of R-PSFs. Finally, the areas of concern are addressed and future research directions are charted.

## PERFORMANCE-SHAPING FACTORS

PSFs can be described as “all these factors such as age, working conditions, team collaboration, mental and physical health, work experience or training which enhance or degrade human performance” (9). Again, researchers are interested mainly in the negative effect of PSFs on human performance.

PSFs can be categorized as internal and external according to their characteristics. The latter are related to situation characteristics, job and task characteristics, or environmental circumstances whereas internal PSFs refer to individual characteristics (10). The internal–external type of PSF categorization is widely accepted and often applied by researchers (11).

A recent study divides PSFs into direct and indirect, with the former defined as those that “can be measured directly, whereby there is a one-to-one relationship between the magnitude of the PSF and that which is measured” (9). Conversely, indirect PSFs are defined as those that “cannot be measured directly, whereby the magnitude of the PSF can only be determined multivariately or subjectively” (9). Although the distinction between direct and indirect PSFs is useful, without clear criteria to distinguish between them, a researcher’s opinion may vary from case to case. In addition, verifying the validity of PSFs is a matter of concern for researchers (9). Therefore, the use

of direct and indirect PSFs may be inefficient and pose considerable problems for researchers.

Several PSF taxonomies are addressed in the literature (3, 12–14), including THERP, SLIM, PHECA, HEART, and CREAM (15). In general, PSF taxonomies are applied to a specific domain and for a predefined purpose. The different taxonomies can be divided into those that consist of a detailed set of PSFs and those that include a more generic set. However, detailed and generic taxonomies have considerable similarities and include similar PSFs regardless of their domain, for development and application.

Detailed PSF taxonomies comprise factors such as lighting, temperature, training, or stress. By contrast, generic taxonomies include factors such as working conditions, organizational factors, or the quality of procedures. Because the set of the possible PSFs is limited, there is an obvious overlap between the generic and detailed taxonomies, that is, generic taxonomies are expressed by several categories, which contain individual detailed PSFs. Another issue relates to the definitions given to PSFs in different taxonomies. Researchers describe identical or similar generic factors with different names or characteristics. For instance, working conditions can be described as the nature of the physical conditions such as ambient light, noise, or temperature (3). This definition, however, is similar to the environmental conditions according to another taxonomy [e.g., Kim and Jung (15)].

The effect of individual PSFs on human performance is a matter of concern for researchers because the PSFs do not affect humans equally. However, the identification of the exact influence of each factor on performance is not an explicit process; for example, for Task A, lighting or training may influence human performance more than communication, whereas the reverse may be the case for Task B. Few human reliability assessment techniques take this issue into account (16). Certain techniques, such as CREAM (3) and SLIM (17), consider the differences in the influence of PSFs on humans and use simple equations to estimate this (3, 18), yet whether they provide reliable results is not justified (5).

Finally, the relationship between PSFs and time poses considerable problems because many PSFs are referred to as “dynamic” variables, changing continuously even while a task is executed, for example, weather conditions or fatigue. Because estimating these changes is too complex, most techniques neither account for this estimation nor analyze the changes separately for every condition. In this paper PSFs that do not change while a task is executed—for example, experience or safety culture—are referred to as “static.”

## R-PSF TAXONOMY

Given the limitations of PSFs discussed in the previous section (direct–indirect; dynamic–static; individual influence of a PSF on human performance), a new, simple, and detailed PSF taxonomy is proposed here to enable researchers (regardless of experience in the field of human factors) to better examine and study human performance. Because the number of PSFs is finite, the taxonomy is related to the existing factors.

With regard to the railway industry, several approaches have been applied, yet only a few of them have introduced new PSF taxonomies compared with corresponding taxonomies from other domains. However, none of these has managed to overcome the limitations described in the previous section (19). Therefore, it is important to develop a PSF taxonomy primarily for the railway domain that accounts for current weaknesses. The new taxonomy, introduced earlier, is referred to in this paper as “railway performance-shaping factors.”

The R-PSF taxonomy has been derived from an extensive literature review in the field of transportation as well as from other domains, such as the nuclear, healthcare, and offshore energy exploration and production industries. The R-PSF taxonomy aims to

- Identify, define, and categorize, on the basis of their common characteristics, those PSFs that influence human performance on railway operations;
- Assess (weight) PSFs according to each operator’s duties (e.g., train driver, signaller, and controller) to propose mitigation strategies;
- Investigate and measure interdependencies between PSFs; and
- Account for the distinction between dynamic and static PSFs.

Although a number of taxonomies for the railway industry were identified in the literature, these taxonomies were either incomplete (20) or strongly oriented toward one type of operator (8). Therefore there is a need for a new taxonomy focused on the railway industry, rather than for adapting an existing one.

The R-PSF taxonomy was developed on the basis of the duties of the railway employees to provide researchers, experienced or younger, as well as operators and safety specialists in the field of human factors, with a simple and comprehensive tool. It defines the PSFs in detail and provides an example for each PSF to avoid potential misunderstandings among researchers. The new taxonomy bridges the identified limitations as presented in the previous section because it

- Clearly defines the PSFs,
- Distinguishes PSFs as dynamic and static,
- Identifies the interactions between PSFs,
- Identifies and determines PSFs that are related to operators’ interactions, and
- Is concerned with transferability and developed not only on the basis of rail operators’ duties but also on the duties of other railway employees, such as maintenance or construction personnel.

Furthermore, to capture more PSFs, taxonomies from domains other than transportation are taken into account.

## Structure of R-PSF Taxonomy

A detailed literature review was performed to identify the factors that affect human performance on railway operations. The literature shows that the number of factors is limited regardless of the task or the operational procedures (16). However, the influence of an individual PSF on human performance is directly related to the task or to the operational procedures (16). Therefore, this paper focuses on identifying the factors that have an effect on the performance of train drivers, signallers, and controllers.

The review considered previous studies in the transport domain (19–26) as well as in other domains such as the nuclear power plant, chemical power plant, and health care industries to cover more relevant entities (3, 9, 15, 16, 27, 28). Thirteen taxonomies that are well-known, well-documented, and widely implemented were selected (5, 16, 17). The literature also shows that beyond a certain point, the existing taxonomies have a tendency to be repeated (2). Therefore, although the selected taxonomies may not be exhaustive, it can be reasonably argued that they contain the majority of the factors that influence human performance in every working environment.

The selected taxonomies are THERP (16), ATHEANA (29), SLIM (16), SPAR-H (30), CREAM (3), and HRMS (16) from the nuclear industry; CARA (24), rail-specific human factors for railways (8), TRACER (31), and HERMES (32) from the transport domain; HEPI (33) from the offshore domain; and the generic HEART (16).

Two hundred forty-eight PSFs were initially identified, and a mapping among them was performed on the basis of their definitions to narrow down the list of factors. Some of the most frequently identified factors are training, distraction, communication, quality of procedures, work experience, time pressure, task complexity, fatigue, or stress. Finally 45 PSFs were derived for the R-PSF taxonomy considering the duties of the train drivers, signallers, and controllers as specified in railway operational manuals (34–36).

For the R-PSF structure, the following approach was adopted:

1. A literature review was conducted to identify and analyze the underlying concept of any previous PSF models (20).
2. Interviews with subject matter experts, that is, railway stakeholders and academics in the field of human factors, were undertaken to determine the interconnections between the PSFs.
3. The individual PSFs were categorized into seven categories according to their definitions and attributes.
4. A simplified model was adopted and modified to provide analysts with a clear and comprehensive structure (20).
5. The final structure was verified and validated by the subject matter experts.

Figure 2 shows the R-PSF structure as well as a sample of the railway PSFs. The model illustrates the interactions between the operator, executed task, and railway PSFs. Figure 2 illustrates the structure of a taxonomy developed for the railway industry, and its structure can be described as generic. Therefore, it is claimed here that the structure of this R-PSF taxonomy, as well as the included R-PSFs, can be used in any other transport mode or industry, although the individual PSFs might change according to the attributes and features of the industry in which they will be implemented.

The PSFs are divided into seven categories. Two of them contain the dynamic factors, that is, those that are strongly related to the precise moment of the operation, and the remaining five contain the static factors. The categories are described as follows:

1. Personal factors, for example, dynamic or static, characterize every individual. For instance, an operator's level of stress for a particular situation A is unique and can be different from the corresponding level of stress of any other operator.

2. Task factors characterize features of the executed task such as its complexity.

3. Team factors influence the operator as a member of a team, for example, communication during shunting.

4. Organizational factors have a significant effect on human performance at the workplace, as affected by the characteristics of the organization for which people work (26). For instance, if a train driver fails to observe a signal because of fatigue, it is an issue that is related to personal factors. However, if the driver is tired because of long consecutive shifts, then the issue is linked to the organization because it is responsible for shift patterns.

5. System factors describe factors such as the quality and type of the equipment or the working conditions in regard to the working environment.

6. Finally, ambient factors include weather conditions at the moment of the operation.

The PSF "time pressure (time to respond)" and "workload" are considered as dynamic factors only in the case of an emergency or unexpected situation.

It is clear that the factors have either a direct or an indirect effect on each other. For instance, adverse weather conditions may directly affect operators' capability to control the train, whereas some organizational factors such as the level of an organization's safety culture have an indirect effect on the operator's performance. The interdependencies among R-PSFs are an issue of great concern and are currently being investigated.

A detailed definition and an example for each one of the R-PSFs is provided to avoid potential misunderstanding among researchers with different backgrounds or biases (37).

Overall, the R-PSF taxonomy provides a framework that can be applied to study other railway employees' performance as well, for example, maintenance personnel or station personnel. However, researchers should be aware of any changes in the context of the duties.

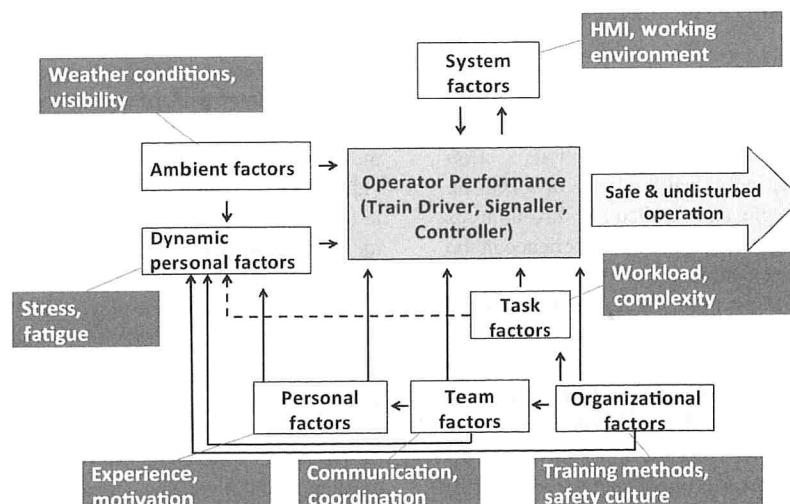


FIGURE 2 The R-PSF structure and a sample of PSFs per category (HMI = human machine interface).

## Validation of R-PSF Taxonomy

To validate the taxonomy, that is, crosscheck, verify, and confirm the literature findings with real data, 179 accident and incident railway reports were analyzed. The reports have been collected from several railway stakeholders: 95 reports (63 accidents, 32 incidents) from several European national investigation bodies, derived from the European Railway Agency Database of Interoperability and Safety; 69 reports from a European railway operator, the Swiss Federal Railways (SBB); and 15 reports (8 accidents, 7 incidents) from the Australian Transport Safety Bureau. The analysis of the reports was conducted in three stages and is presented in the next section. Stage 1 involved the identification of the characteristics of the most frequent occurrences, including the type of trains involved and the immediate cause that led to that occurrence. Then PSFs that contributed to the occurrence were identified, and finally an expert's elicitation assessment was conducted in collaboration with SBB. The purpose of this assessment was to rank the influence of the PSFs on human performance for four specific railway operations, as well as to identify any missing PSFs that according to operators' experience play an important role in their performance.

The reports contain information on the following: the type of railway, for example, regional passenger train; the occurrence type, for example, accident or incident; the associated event, for example, train derailment; the location and time of the accident; the immediate cause of the accident, for example, train unable to stop; the causal factor of the accident, for example, train driver falls asleep; what PSFs played a role in that occurrence; and the severity of its consequences. The sources of the reports represent most of the situations experienced in railway operations, for example, different types of networks, varying geography (Australia and Switzerland), different regulations, and different staffing (two drivers on the Australian trains instead of one on European trains). Therefore, it can be argued that the sampled data are representative of the population of interest.

As explained in the previous sections, the definitions provided by the R-PSF taxonomy were used to define the PSFs that occur in railway operations. Subsequently the Pareto principle, which shows that the majority of the results come from a minority of inputs on an 80–20 basis, was used to determine those PSFs that most frequently occur, as can be seen in Table 1. Therefore, on the basis of the Pareto principle, 18 PSFs are responsible for more than 80% of the occurrences (from the 179 reports), with respect to immediate causes, as shown in the next section.

## Assessment of R-PSFs

To assess the R-PSF taxonomy, a study was conducted in collaboration with a widely recognized European railway organization, SBB. The main reasons for selecting SBB were the organization's excellent reputation for safety, easy and complete access to data and reports, availability and accessibility of SBB personnel, significant investment in human factors, and the complexity of its network. Although small, the SBB network is complex and dense. Its complexity is characterized by 804 stations in a 3,139-km-long network, almost 347 million passengers, and 50 million tons of goods transported every year (2010 figures, personal communication with Andreas Hönger). The SBB network's extremely high punctuality percentage, combined with the lack of a dedicated freight train line and the multiple maintenance procedures on the network, underpins the characteristics of its complexity.

In this study, 18 SBB employees participated, including two risk analysts, three passenger train drivers, six freight train drivers, and six signalers. All the employees had at least 5 years of experience in their domain, and some had been SBB employees for almost 30 years. Currently train drivers are dedicated to either passenger or freight trains.

According to the results from the analysis of the reports, four different scenarios were designed in accordance with their frequency and the severity of consequences. From the collected reports no variation was identified in regard to the frequency of the occurrences. The SBB employees were asked to assess the PSFs presented in Table 1 for each of the scenarios and to identify potential interdependencies among those PSFs. Three of the scenarios are related to train drivers, with the remaining scenario related to signalers, as follows:

Scenario 1. Long-distance passenger train derails due to excessive speed.

Scenario 2. Regional passenger train passes a signal at danger although the driver was aware of it.

Scenario 3. Freight train passes signal at danger because of the wrong interpretation by the driver.

Scenario 4. During a shunting operation a shunter engine collides with a commercial train as the result of a signaller error.

The scenarios contain information, such as the type of train and occurrence; responsible and involved personnel; system information, for example, technical failures; weather conditions; operator's

TABLE 1 Most Frequently Met PSFs on Railway Operations

R-PSF	Category	R-PSF	Category
Distraction–concentration	DP	Experience–familiarity	P
Expectation–routine	DP, T	Time pressure (time to respond)	T
Communication	Te	Fit to work–health	P, O
Safety culture	O	Workload	T
Training–competence	P, O	Visibility	A
Perception	DP	Coordination of work–supervision	Te, O
Quality of procedures	O	Weather conditions	A
HMI quality (e.g., type of locomotive, communication technical failure)	S	Stress	DP
Fatigue (sleep lack, shift pattern)	DP, O	Risk awareness	P

NOTE: DP = dynamic personal; P = personal; T = task; Te = team; O = organizational; A = ambient; S = system.



level of experience; and location, date, and time of the event. The scenarios were refined and checked by SBB risk analysts and subsequently distributed to the participants, who assessed the 18 factors on a scale of 1 to 18, ranging from most to least important for each factor. The assessment of PSFs was conducted on an individual basis, apart from one case in which two participants completed one ranking assessment together. Participants ranked the R-PSFs by employee and primary cause, as included in the four scenarios. The authors' physical presence when 10 out of 18 participants assessed the ranking, together with telephone access to the rest, assured the validity of the responses in relation to any misunderstandings or vagueness experienced by the participants.

To consolidate the results from this study, an extended assessment of the R-PSFs is being conducted in collaboration with most of the organizations that contributed reports, including Network Rail (UK), JBV (Norwegian Rail Administration), and the Dutch Railways (both operators and the network operator); assessment of the scenarios is included.

### Identified Limitations

The collected reports are official documents that have been completed by authorized personnel, from different sources such as national transportation safety boards, independent national investigation bodies, and railway operators. The literature shows that the reporting system in the countries where the reports were completed, despite their limitations, is effective; therefore it could be assumed that they contain reliable information, sufficient to extract accurate outcomes (38).

However, not all of the reports include the minimum necessary information for the validation of R-PSFs. Therefore, an ongoing complementary study is being performed, which monitors the reporting system of different investigation bodies, checks reporting reliability, and tries to establish a weighting factor for their accuracy.

## RESULTS

This section is divided into two parts as follows:

- The first part presents the results of the reports analysis and
- The second illustrates results of the study for the assessment of the railway PSFs.

The results provide a justification of the factors that influence operators in four different railway operations. This justification will be used to develop the HuPeROI, an index that estimates human error probability in railway operations.

In general, the results show that there is a strong relation between the literature findings, the report outcomes, and the results from the pilot study. In addition, the assessment of R-PSFs through the pilot study indicates that most participants' responses confirm the findings from the reports analysis.

Because the results show significant commonalities regardless of the scenario, it is believed that the R-PSF taxonomy and the ranking of R-PSFs can be applied and transferred to several other cases.

### Accident and Incident Analysis

Of the 179 reports, 128 refer to the degraded performance of train drivers, 49 to signalers or controllers, and the remaining two to an error including both driver and signaller.

Figures 3 through 5 present the initial accident and incident analysis. According to the findings in Figure 3, the cases for regional passenger trains (including commuters), freight, and long-distance passenger trains in on-route and shunting operations were investigated. For these types of trains, the immediate causes that led to the occurrences were analyzed, as well as the potential association between the immediate causes and the types of trains.

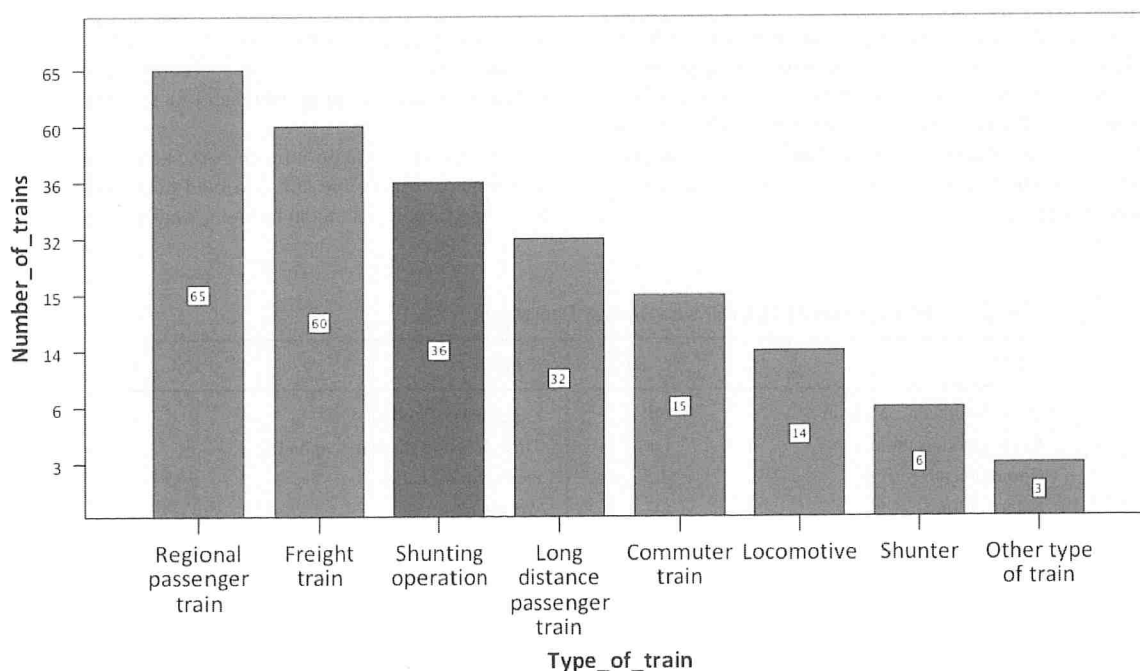


FIGURE 3 Types and number of trains involved in occurrences.

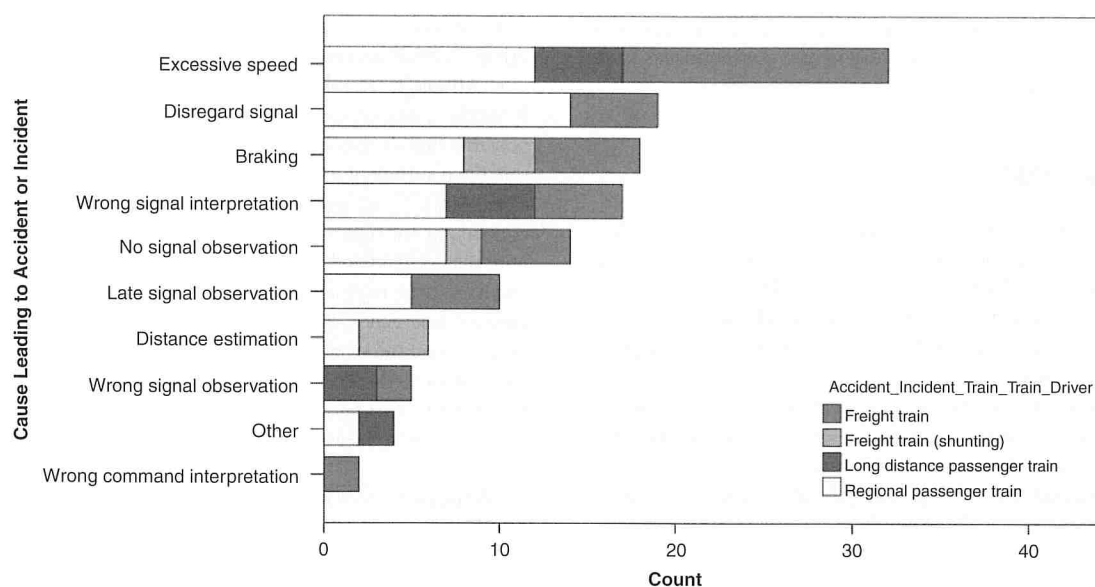


FIGURE 4 Immediate cause leading to accident or incident (train driver).

From Figure 4 it can be seen that excessive speed, signaling, and braking-related errors are the most common. It can also be observed that the three dominant immediate causes are monitored for all types of trains. However, no association was found between "immediate causes–types of trains" as determined by a chi-square test ( $p = .394 > .05$ ), including the events during shunting operations.

Figure 5 illustrates the frequency of PSFs extracted from the reports per type of train driver and signaller or controller (data normalized by the number of relevant reports).

This shows that the factor "distraction–loss of concentration" is the predominant contributing factor that leads to an event, followed by "safety culture." "Expectation–routine" is a major factor for train drivers, whereas "workload" is a significant factor only for signalers. However, "communication between employees" seems to be a significant factor for all types of employees.

An inspection of Figure 5 indicates that not all PSFs were identified in all cases and that they do not contribute equally to operators' performance. In view of that result, two more Pearson chi-square tests were carried out to examine the association of PSFs with operators. One test was executed between signalers–controllers and train drivers, and the other was performed with train drivers only. The first test showed association between "PSF–type of operators" ( $p = .000 < .05$ ) as they are divided into train drivers and signalers–controllers. However, no association was observed between "PSF–types of drivers," as determined by the Pearson chi-square test ( $p = .266 > .05$ ).

On the basis of the severity of the consequences identified from the reports (number of fatalities and material damage), two signaling scenarios and one related to excessive speed were chosen for analysis in the pilot study. As the analysis indicated also that errors by signalers and controllers can occur equally during the on-route

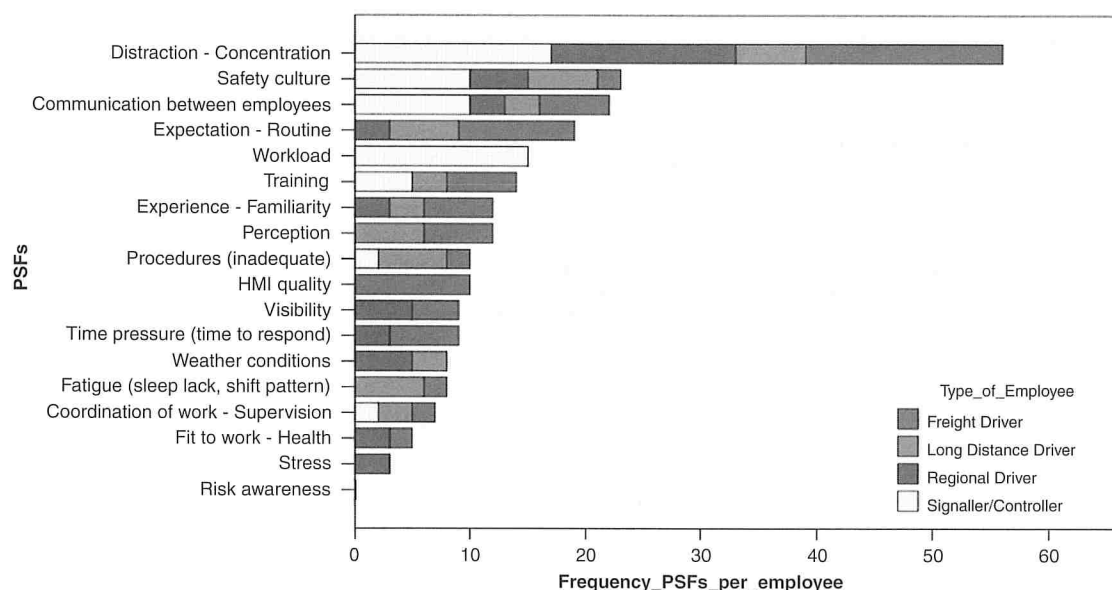


FIGURE 5 Railway PSF frequency per type of employee.

and the shunting procedure, a fourth scenario related to shunting operation was selected because the most severe consequences have been identified for such procedures.

## Assessment of PSFs

The assessment results confirm the findings from accident and incident analysis; selected findings of the assessment results for seven R-PSFs are shown in Table 2. Six of the R-PSFs presented were identified from the reports as the most frequent (Figure 5), and the “stress” PSF illustrates a noticeable contradiction between the reports and assessment results.

Table 2 contains the following information: the four scenarios, the seven R-PSFs and their rankings as identified from the reports, the percentage of SBB employees’ common responses (first figure in column), and the corresponding range of these responses (in parentheses). Consider the example of “distraction–loss of concentration,” which was identified as the most frequent (Ranking 1) R-PSF from the reports. The assessment of SBB employees shows that for the train driver scenarios, more than 75% of employees identify “distraction–loss of concentration” as the most significant R-PSF (Range 1–4) and at least 50% believe the same for the fourth scenario. The rest of the six R-PSFs are illustrated accordingly, in particular, as follows:

- Regardless of the network or the operation, distraction–loss of concentration has been determined to be the most significant R-PSF, which strongly justifies Figure 5.
- “Safety culture” is a significant factor for signalers but not equally important for drivers, which satisfies the results from the analysis, as shown in Figure 5. Although safety culture should be equally important for all types of employees and operations, it is possible that the observed difference is due to the given definition, which includes the cases of “disregarded procedures.” For the selected scenarios, participants believe that it is extremely important for the signalers to comply with the given procedures to keep the operation safe and secure.
- “Communication between employees” was assessed as the main contributor PSF for signalers (shunting operation), which confirms the analysis findings, as illustrated in Figure 5.

- “Workload” is an essential PSF for signalers but not for train drivers, which justifies the results in Figure 5.
- Although “training” was identified as a fairly substantial PSF from the analysis, this was not confirmed by the pilot study. It is possible that the participants assumed or considered that rail operators have sufficient training to execute their tasks.
- “Stress” was identified from the pilot study as a serious PSF contributor for signalers, but not for drivers, which is not consistent with the accident and incident analysis, as shown in Figure 5. On the one hand, with regard to signalers, stress is a major PSF, especially in cases of emergency. It is likely that the reporting system was not providing direct information about this PSF, so analysts have to assume its existence. On the other hand, concerning regional train drivers, it is possible that participants unfamiliar with this type of operation underestimated the contribution of stress.

Although for many PSFs the responses were different, the range of difference was marginal. This suggests that the 1–18 scale used for the assessment should be reconsidered and reduced accordingly.

Finally, because the R-PSFs assessment is strongly related to the scenarios, it is likely that in the case of other scenarios, results may be different.

## CONCLUSIONS

This paper has introduced an R-PSF taxonomy to underpin the subsequent development of the HuPeROI. Because the ultimate aim of this research is to develop the HuPeROI to estimate operator error probabilities in railway operations, it is important to assess and validate those R-PSFs that mainly influence operator performance. Therefore, those events that not only occur most frequently but are also associated with the most severe consequences were defined from the analysis of 179 reports. Four real scenarios were subsequently chosen and experts assessed the R-PSFs that influenced operators in those scenarios. The results were checked and compared with the findings from an analysis of the reports.

The R-PSF taxonomy provides analysts with a detailed list of PSFs that influence operators’ performance; the taxonomy also describes in an appropriate manner the interactions between the

TABLE 2 Assessment of R-PSFs per Scenario

Scenario	PSF Assessment						
	Distraction–Loss of Concentration, 1 <sup>a</sup> [% (range)]	Safety Culture, 2 <sup>a</sup> [% (range)]	Communication Between Employees, 3 <sup>a</sup> [% (range)]	Expectation: Routine, 4 <sup>a</sup> [% (range)]	Workload, 5 <sup>a</sup> [% (range)]	Training, 6 <sup>a</sup> [% (range)]	Stress, 17 <sup>a</sup> [% (range)]
1. Train driver (long-distance passenger train)	86.40 (1–4)	24.63 (1–4)	6.25 (1–4) 75.00 (15–18)	52.994 (1–5)	70.59 (>6)	52.94 (10–14)	35.29 (2–5)
2. Train driver (regional passenger train)	76.47 (1–4)	35.29 (1–4)	11.76 (1–4) 52.94 (15–18)	52.94 (1–5)	70.51 (>6)	58.82 (10–14)	29.41 (2–5)
3. Train driver (freight train)	76.47 (1–4)	11.76 (1–4)	11.76 (1–4) 70.59 (15–18)	52.94 (1–5)	76.47 (>6)	64.71 (10–14)	29.41 (2–5)
4. Signaler	52.94 (1–4)	52.94 (1–4)	52.94 (1–4) 17.65 (15–18)	47.06 (1–5)	70.59 (3–6)	52.94 (10–14)	70.59 (2–5)

<sup>a</sup>Ranking as identified from reports.

operator, the executed task, and the PSFs. In addition, the taxonomy distinguishes the R-PSFs with regard to their dependency on time as dynamic and static, a distinction which is essential for the development of HuPeROI.

To assess the R-PSF taxonomy, an analysis of railway accident and incident reports was performed together with a pilot study. The results are in general agreement with those of previous similar studies (39). However, the findings of this paper add to previous studies because the findings provide evidence for the association between PSFs, immediate causes that lead to an event, types of trains, and types of operators.

To enhance this study, a review of 120 additional rail accident and incident reports is ongoing. An additional study, scheduled in collaboration with another rail organization, will seek to justify to a greater extent the R-PSF assessment. Finally, the interdependencies between the railway PSFs will be defined.

## ACKNOWLEDGMENTS

The authors are grateful for the active participation of the Swiss Federal Railways (SBB) in this study. In particular, the authors thank Andreas Hönger for all his help and support and all SBB employees who participated and commented on this research. Support from the Lloyd's Register Educational Trust in undertaking this study is also gratefully acknowledged.

## REFERENCES

1. Dhillon, B. S. *Human Reliability and Error in Transportation Systems*. Springer, London, 2007.
2. Wilson, J. R., T. Farrington-Darby, G. Cox, R. Bye, and G. R. J. Hockey. The Railway as a Socio-Technical System: Human Factors at the Heart of Successful Rail Engineering. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, Vol. 221, 2007, pp. 101–115.
3. Hollnagel, E. *Cognitive Reliability and Error Analysis Method*. Elsevier, New York, 1998.
4. Priestley, K., and G. Lee. Human Factors in Railway Operations. Presented at International Conference on Challenges for Railway Transportation in Information Age, Hong Kong, China, 2008.
5. Bell, J., and J. Holroyd. *Review of Human Reliability Assessment Methods*. Health & Safety Laboratory, Norwich, United Kingdom, 2009.
6. Maurino, D. ICAO Supports Proactive Approach to Managing Human Factors Issues Related to Advanced Technology. *ICAO Journal*, 1998.
7. Evans, A. W. Fatal Train Accidents on Europe's Railways: 1980–2009. *Accident Analysis and Prevention*, Vol. 43, 2011, p. 391–401.
8. RSSB. *Rail-Specific HRA Technique for Driving Tasks*. Rail Safety and Standards Board, London, 2004.
9. Boring, R. L., C. D. Griffith, and J. C. Joe. The Measure of Human Error: Direct and Indirect Performance Shaping Factors. Presented at Joint 8th IEEE Conference on Human Factors and Power Plants/13th Conference on Human Performance, Root Cause and Trending (IEEE HFPP & HPRCT), 2007; Idaho National Laboratory, U.S. Nuclear Regulatory Commission, 2007.
10. Kumamoto, H., and E. J. Henley. *Probabilistic Risk Assessment and Management for Engineers and Scientists*, 2nd ed. IEEE Press, New York, 1996.
11. Sasou, K., and J. Reason. Team Errors: Definition and Taxonomy. *Reliability Engineering and System Safety*, Vol. 65, 1999.
12. Rasmussen, J. Human Errors. A Taxonomy for Describing Human Malfunction in Industrial Installations. *Journal of Occupational Accidents*, Vol. 4, No. 2–4, 1982, pp. 311–333.
13. Swain, A. D., and H. E. Guttman. *Handbook of Human Reliability Analysis with Emphasis on Nuclear Power Plant Applications*. U.S. Nuclear Regulatory Commission, 1983.
14. Williams, J. C. A Data-Based Method for Assessing and Reducing Human Error to Improve Operational Procedures. Presented at 4th IEEE Conference: Human Factors and Power Plants, Monterey, Calif., 1988.
15. Kim, J. W., and W. Jung. A Taxonomy of Performance Influencing Factors for Human Reliability Analysis of Emergency Tasks. *Journal of Loss Prevention in the Process Industries*, Vol. 16, 2003, pp. 479–495.
16. Kirwan, B. *A Guide to Practical Human Reliability Assessment*. Taylor & Francis Ltd., Bristol, Pa., 1994.
17. Forester, J., A. Kolaczowski, E. Lois, and D. Kelly. *Evaluation of Human Reliability Analysis Methods Against Good Practices*. NUREG-1842. U.S. Nuclear Regulatory Commission, 2006.
18. Embrey, D. E., P. Humpherys, E. A. Rosa, B. Kirwan, and K. Rea. *SLIM-MAUD: An Approach to Assessing Human Error Probabilities Using Structured Expert Judgement*. U.S. Nuclear Regulatory Commission, Washington, D.C., 1984.
19. Cacciabue, P. C. Human Error Risk Management Methodology for Safety Audit of a Large Railway Organisation. *Applied Ergonomics*, Vol. 36, pp. 709–718.
20. Hammerl, M., and F. Vanderhaegen. Human Factors in Railway System Safety Analysis Process. Presented at 3rd International Human Factors Conference, Lille, France, 2009.
21. Hamilton, I. W., and T. Clarke. Driver Performance Modelling and Its Practical Application to Railway Safety. *Applied Ergonomics*, Vol. 36, 2005, pp. 661–670.
22. *Overview of HRA Methods—Farandole Project*. European Organisation for the Safety of Air Navigation (EUROCONTROL), Brussels, Belgium, 2007.
23. *Understanding Human Factors: A Guide for the Railway Industry*. Rail Safety and Standards Board, London, 2004.
24. Kirwan, B., and H. W. Gibson. CARA: A Human Reliability Assessment Tool for Air Traffic Safety Management—Technical Basis and Preliminary Architecture. In *15th Safety-Critical Systems Symposium*, Springer, Bristol, United Kingdom, 2007.
25. Baysari, M. T., A. S. McIntosh, and J. R. Wilson. Understanding the Human Factors Contribution to Railway Accidents and Incidents in Australia. *Accident Analysis and Prevention*, Vol. 40, 2008, pp. 1750–1757.
26. Reason, J. *Managing the Risks of Organizational Accidents*. Ashgate Publishing Ltd., Farnham, United Kingdom, 1997.
27. Trucco, P., and M. C. Leva. A Probabilistic Cognitive Simulator for HRA Studies (PROCOS). *Reliability Engineering and System Safety*, Vol. 92, 2007, pp. 1117–1130.
28. Lyons, M., S. Adams, M. Woloshynowych, and C. A. Vincent. Human Reliability Analysis in Healthcare: A Review of Techniques. *International Journal of Risk and Safety in Medicine*, Vol. 16, 2004, pp. 223–237.
29. Forester, J., A. Kolaczowski, S. Cooper, D. Bley, and E. Lois. *ATHE-ANA User's Guide*. U.S. Nuclear Regulatory Commission, 2007.
30. Gertman, D., H. Blackman, J. Marble, J. Byers, and C. Smith. *The SPAR-H Human Reliability Analysis Method*. Idaho National Laboratory, U.S. Nuclear Regulatory Commission, 2004.
31. Shorrock, S. T., and B. Kirwan. Development and Application of a Human Error Identification Tool for Air Traffic Control. *Applied Ergonomics*, Vol. 33, 2002, pp. 319–336.
32. Cacciabue, P. C. Human Error Risk Management for Engineering Systems: A Methodology for Design, Safety Assessment, Accident Investigation and Training. *Reliability Engineering and System Safety*, Vol. 83, 2004, pp. 229–240.
33. Khan, F. I., P. R. Amyotte, and D. G. DiMattia. HEPI: A New Tool for Human Error Probability Calculation for Offshore Operation. *Safety Science*, Vol. 44, 2006, pp. 313–334.
34. *National Operations Centre Controller Job Description*. Network Rail, London, 2008.
35. *National Operations Signaller Job Description*. Network Rail, London, 2008.
36. *National Operations Train Driver Job Description*. Network Rail, London, 2008.
37. Kyriakidis, M. *Railway Performance Shaping Factors—Definitions and Examples*. Imperial College London, 2011.
38. Valonen, K. Investigation of Rail Accidents—International Comparison. In *Mechanical Engineering*, University of Technology, Helsinki, Finland, 2000.
39. Fisher, K. *Signalfälle. Eine arbeitspsychologische Ursachenanalyse*. Fachhochschule Nordwestschweiz FHNW, Olten, Switzerland, 2007.

*The Railroad Operational Safety Committee peer-reviewed this paper.*

# Analysis of Causes of Major Train Derailment and Their Effect on Accident Rates

Xiang Liu, M. Rapik Saat, and Christopher P. L. Barkan

Analysis of the causes of train accidents is critical for rational allocation of resources to reduce accident occurrence in the most cost-effective manner possible. Train derailment data from the FRA rail equipment accident database for the interval 2001 to 2010 were analyzed for each track type, with accounting for frequency of occurrence by cause and number of cars derailed. Statistical analyses were conducted to examine the effects of accident cause, type of track, and derailment speed. The analysis showed that broken rails or welds were the leading derailment cause on main, yard, and siding tracks. By contrast to accident causes on main tracks, bearing failures and broken wheels were not among the top accident causes on yard or siding tracks. Instead, human factor–related causes such as improper use of switches and violation of switching rules were more prevalent. In all speed ranges, broken rails or welds were the leading cause of derailments; however, the relative frequency of the next most common accident types differed substantially for lower-versus higher-speed derailments. In general, at derailment speeds below 10 mph, certain track and human factor causes—such as improper train handling, braking operations, and improper use of switches—dominated. At derailment speeds above 25 mph, those causes were nearly absent and were replaced by equipment causes, such as bearing failure, broken wheel, and axle and journal defects. These results represent the first step in a systematic process of quantitative risk analysis of railroad freight train safety, with an ultimate objective of optimizing safety improvement and more cost-effective risk management.

Train accidents cause damage to infrastructure and rolling stock as well as service disruptions, and may cause casualties and harm the environment. Accordingly, improving train operating safety has long been a high priority of the rail industry and the government. Train accidents occur as a result of many different causes; however, some are much more prevalent than others. Furthermore, the frequency and severity of accidents also varies widely, depending on the particular accident cause (1–3). Efficient allocation of resources to prevent accidents in the most cost-effective manner possible requires understanding which factors account for the greatest risk, and under which circumstances. Assessment of the benefits and costs of strategies to mitigate each accident cause can then be evaluated and resources allocated so that the greatest safety improvement can be

achieved for the level of investment available. This paper presents statistical results representing the first step in a systematic process of quantitative risk analysis and risk management for railroad freight train safety.

## APPROACH

The approach taken in this research is to conduct detailed analysis of the train accident data supplied by the railroads to FRA of the U.S. Department of Transportation. FRA regularly publishes statistical summaries of these data; however, the results are generally presented at a highly aggregated level. Further insights are possible by analyzing the results in more detail and considering other statistical approaches. In addition, there are various metrics that can be used to assess train safety. The effectiveness of specific risk reduction strategies needs to be understood when the cost-effectiveness of research, development, and implementation of new strategies is considered. Consequently, in the final section of this paper a preliminary sensitivity analysis of several groups of accident causes is conducted to understand how changes in practice or failure prevention technology might affect the overall accident rate. The results enable objective comparison of different approaches that could be used to inform decision making by industry and government concerning which research, development, or implementation strategies to invest in.

## DATA SOURCES AND ANALYSIS

FRA requires railroads to submit detailed reports of all significant accidents or incidents associated with railroad train operation. It is useful to review briefly the FRA databases in the larger context of railroad safety and analysis, including how the databases relate to one another and the hierarchical organization of the train accident database, which is the subject of the research described in this paper. These databases can be considered at increasing levels of detail as follows: type of incident (corresponding to particular FRA databases) and, within the database on train accidents, by track type, accident type, and accident cause.

## FRA DATABASES

FRA maintains three major databases, each related to a different aspect of train operating safety: train accidents, employee casualties, and railroad and highway grade crossing incidents. A particular

---

Rail Transportation and Engineering Center, Department of Civil and Environmental Engineering, University of Illinois at Urbana–Champaign, 205 North Mathews Avenue, Urbana, IL 61801. Corresponding author: X. Liu, liu94@illinois.edu.

*Transportation Research Record: Journal of the Transportation Research Board*, No. 2289, Transportation Research Board of the National Academies, Washington, D.C., 2012, pp. 154–163.  
DOI: 10.3141/2289-20



reportable event may require that reports be submitted to any or all of these, alone or in combination, depending on the circumstances. Of principal interest for the research described in this study is the first database describing the circumstances, physical characteristics, and other information related to damage to rolling stock and infrastructure. Within this database the interest lies in the type of track—main, siding, yard, or industry—that an accident occurred on. At the next level down, interest is focused on the type of accident, that is, derailment, collision, or various other types. Finally, within each of these track and accident types the particular cause of the accident and other circumstances, notably derailment speed, are the focus of interest.

### Event Category and Corresponding FRA Database

The Rail Equipment Accident/Incident Report (REAIR) form (FRA F 6180.54) is used by railroads to report all accidents that exceed a monetary threshold of damages to infrastructure and rolling stock. [The form accounts for damage to on-track equipment, signals, track, track structures, and roadbed. The reporting threshold is periodically adjusted for inflation and increased from \$7,700 in 2006 to \$9,400 in 2011 (4).] FRA compiles these reports into the rail equipment accident (REA) database, which records rail equipment accident data dating back to 1975. In addition to the REAIR, the Highway–Rail Grade Crossing Accident/Incident Report (FRA F 6180.57) and Death, Injury, or Occupational Illness Summary (FRA F 6180.55a) are the other two principal eponymous railroad accident and incident reporting forms. A single accident may require more than one report, depending on its circumstances. For example, if a train accident occurs that results in damages to track and equipment exceeding the threshold, an FRA F 6180.54 report must be submitted, and if the accident involved a highway user at a highway–rail crossing, regardless of impact, a Form FRA F 6180.57 must also be completed. All casualties resulting from a reportable rail equipment accident, in addition to being recorded on Form FRA F 6180.54, must also be reported individually on Form FRA F 6180.55a (4). This study used data exclusively from the FRA REA database. Depending on the nature of one's interest in train accident analysis, additional useful information may be found in the other databases and these databases can be linked to pursue additional questions not possible with a particular database alone (5).

### FRA Rail Equipment Accident Database

The FRA REA database records railroad, accident type, location, accident cause, severity, and other information important for accident analysis and prevention. This paper focuses on Class I freight railroads (operating revenue exceeding \$378.8 million in 2009), which account for approximately 68% of U.S. railroad route miles, 97% of total ton-miles transported, and 94% of the total rail freight revenue (6). In addition to analysis of the number of freight trains derailed due to various causes, consideration of statistics on the number of cars derailed and the circumstances of their derailment is necessary because accident severity varies among different accident causes. To understand the effect of various derailment prevention strategies, first there is a need to quantify how much different accident causes contribute to derailment risk and also how accident characteristics affect the risk.

### ACCIDENT BY TRACK TYPE

Four types of tracks are recorded in the FRA REA database—main, siding, yard, and industry tracks. These track types are used for different operational functions and consequently have different associated accident types, causes, and consequences. Train accidents are categorized into derailment, collision, highway–rail grade crossing accident, and several other less frequent types. When there is more than one type of accident, the type of accident that occurred first would be designated for all reports related to it. For example, a derailment caused by a collision would be classified as a “collision.” Highway–rail grade crossing accidents in the REA database include only those that occur at the highway–rail interface and involve at least one highway user (4).

FRA-reportable freight train accident data for Class I railroads for the period 2001 to 2010 were compiled to show the number of FRA-reportable accidents, the average number of cars derailed per accident, and the total number of cars derailed by accident type and track type (Table 1). Train derailment was the most common type of accident on each track type, and train collision was the least frequent (excluding highway–rail grade crossing accidents on siding, yard, and industry tracks). Ninety-eight percent of highway–rail grade crossing accidents occurred on main track and accounted for 20% of all types of Class I main-line train accidents. By definition, these accidents exceeded the FRA reporting threshold for damages, but often did not result in a derailment (5). Accident severity is defined in this study as the number of cars derailed per accident and varies by track type and accident type. Train derailments on main and siding tracks had a greater average accident severity than did other types of accidents and tracks. Highway–rail grade crossing accidents had fewer cars derailed per accident because many reportable highway–rail grade crossing accidents resulted in no derailment (5).

Total number of cars derailed accounts for accident frequency and severity. The majority of cars derailed on Class I freight railroads were derailed as a result of train derailments. Derailments on main and siding tracks accounted for 65% of freight train accidents and correspondingly 87% of the cars derailed on all types of track.

It is evident that the distribution of accident types varied by track type. For example, 98% of highway–rail grade crossing accidents occurred on main tracks, whereas far fewer occurred on yard tracks. A chi-square test was used to examine the association between track type (main, siding, yard, and industry) and accident type (derailment, collision, highway–rail grade crossing accident, and other) by accident frequency. The chi-squared test showed that the accident frequency distributions of different accident types varied by track type ( $\chi^2 = 1,054$ ,  $df = 9$ ,  $P < .01$ ). This result was significant even when only derailment and collision were included in the analysis ( $\chi^2 = 68$ ,  $df = 3$ ,  $P < .01$ ). The association between accident type and track type implies that different track types have different accident cause distributions, which will be discussed in the following sections. Train collisions and highway–rail grade crossing accidents have been analyzed in other recent studies, so this research focused on train derailments (5, 7, 8).

### TRAIN ACCIDENT CAUSE

FRA train accident cause codes are hierarchically organized and categorized into major cause groups—track, equipment, human factors, signal, and miscellaneous. Within each of these major cause groups, FRA organizes individual cause codes into subgroups of

**TABLE 1** Accident Frequency, Accident Severity, and Car Derailment by Accident Type and Track Type, Class I Freight Railroads, 2001–2010

	Accident Type				
Track Type	Derailment	Collision	Highway–Rail	Other	All Accident Types
Number of Freight Train Accidents					
Main	4,439	302	1,343	590	6,674
Yard	2,848	355	12	378	3,593
Siding	436	23	4	40	503
Industry	369	21	6	49	445
All	8,092	701	1,365	1,057	11,215
Average Number of Cars Derailed per Accident					
Main	8.4	3.3	0.5	1.0	5.9
Yard	4.7	1.5	0.8	1.4	4.0
Siding	5.7	3.7	0.0	1.2	5.2
Industry	4.3	1.0	1.3	0.5	3.7
All	6.8	2.3	0.5	1.1	5.2
Total Number of Cars Derailed					
Main	37,456	989	609	580	39,634
Yard	13,363	527	9	511	14,410
Siding	2,477	85	0	47	2,609
Industry	1,593	22	8	23	1,646
All	54,889	1,623	626	1,161	58,299

related causes, such as roadbed and track geometry, within the track group and similar subgroups within the other major cause groups. A variation on the FRA subgroups developed by Arthur D. Little, in which similar cause codes were combined into groups on the basis of expert opinion, was used (9, 10). The Arthur D. Little groupings are similar to FRA's subgroups but are more fine-grained, thereby allowing greater resolution for certain causes. For example, FRA combines broken rails, joint bars, and rail anchors in the same subgroup, whereas the Arthur D. Little grouping distinguishes between broken rail and joint bar defects. These groups were used to analyze cause-specific derailment frequency and severity. The cause groups are ranked in descending order by number of derailments and total number of cars derailed, respectively. The former metric pertains to derailment frequency, whereas the latter accounts for derailment frequency and severity. Different ranking methods may lead to different safety improvement prioritization decisions.

## TRAIN DERAILMENTS

Derailments are the most common type of train accident in the United States, and preventing them has long been a focus of the rail industry and the government (1–3, 9–25). Most previous studies have focused on main-line derailments, with less research published on yard and siding derailments. The derailment-cause distribution on main lines differs from distributions on yard or siding tracks, in part because of the different nature of operations in these two settings. Understanding the top causes affecting train derailment occurrence and number of cars derailed on different tracks provides additional insight into the development, evaluation, prioritization, and implementation of accident prevention strategies given a specific set of operating conditions.

## Derailments on Main Tracks

Although serious incidents can and do occur on yard and siding tracks, the focus of this research is on main-line derailments because of the higher speeds and longer consists typical of main-line operation. The greater mass and speed mean that the force and potential impact in regard to property damage, casualties, and environmental effects are all correspondingly greater. An analysis of derailment causes was conducted to compare the relationship between frequency and severity by derailment cause. Accounting for severity is important because derailments in which more cars are involved are likely to be more damaging and more costly, have a greater likelihood of involving a hazardous materials car if any are in the consist, and if derailed they are more likely to suffer a release (1). In addition to type of track, accident severity, as measured by number of cars derailed per accident, also varies by accident cause (Table 2). Accident severity is affected by a variety of factors, including train length (2, 16, 20), derailment speed (1–3, 13, 16, 17, 20, 25), point of derailment (POD) (the position of the first car in the train that is derailed) (2, 16), and other factors (11, 12). A number of studies have investigated the parametric relationships between accident cause and certain contributing factors affecting train derailment severity (2, 13, 16). The number of derailments and total number of cars derailed are directly related to train derailment rate and car derailment rate, respectively. The former represents the likelihood that a train is involved in a derailment and the latter the likelihood of an individual car derailing. Both rates are useful in risk assessment depending on the question being addressed (2, 3, 9, 13, 17, 19–25). The importance of either statistic can be used to rank the importance of a particular accident cause, and these were used to investigate the association between the two ranking methods (Table 2). A Spearman's rank correlation test showed that the two ranking methods were significantly related

TABLE 2 Derailment Frequency and Severity by Accident Cause on Class I Main Lines, Sorted by Frequency

Cause Group	Description	Derailments		Cars Derailed		Average Number of Cars Derailed per Derailment
		Number	Percentage	Number	Percentage	
08T	Broken rails or welds	665	15.3	8,512	22.7	12.8
04T	Track geometry (excluding wide gauge)	317	7.3	2,057	5.5	6.5
10E	Bearing failure (car)	257	5.9	1,739	4.6	6.8
12E	Broken wheels (car)	226	5.2	1,457	3.9	6.4
09H	Train handling (excluding brakes)	201	4.6	1,553	4.1	7.7
03T	Wide gauge	169	3.9	1,729	4.6	10.2
01M	Obstructions	153	3.5	1,822	4.9	11.9
05T	Buckled track	149	3.4	1,891	5.0	12.7
04M	Track–train interaction	149	3.4	1,110	3.0	7.4
11E	Other axle or journal defects (car)	144	3.3	1,157	3.1	8.0
03M	Lading problems	134	3.1	791	2.1	5.9
07E	Coupler defects (car)	133	3.1	771	2.1	5.8
13E	Other wheel defects (car)	129	3.0	668	1.8	5.2
09E	Sidebearing, suspension defects (car)	126	2.9	816	2.2	6.5
10T	Turnout defects: switches	118	2.7	601	1.6	5.1
11H	Use of switches	104	2.4	407	1.1	3.9
06E	Centerplate or carbody defects (car)	98	2.3	507	1.4	5.2
01H	Brake operation (main line)	95	2.2	881	2.4	9.3
12T	Miscellaneous track and structure defects	80	1.8	687	1.8	8.6
01T	Roadbed defects	67	1.5	665	1.8	9.9
07T	Joint bar defects	66	1.5	1,040	2.8	15.8
10H	Train speed	61	1.4	403	1.1	6.6
09T	Other rail and joint defects	56	1.3	1,132	3.0	20.2
19E	Stiff truck (car)	55	1.3	365	1.0	6.6
05M	Other miscellaneous	54	1.2	422	1.1	7.8
15E	Locomotive trucks, bearings, wheels	50	1.1	177	0.5	3.5
18E	All other car defects	47	1.1	248	0.7	5.3
06T	Rail defects at bolted joint	46	1.1	927	2.5	20.2
12H	Miscellaneous human factors	44	1.0	377	1.0	8.6
02T	Nontraffic, weather causes	43	1.0	331	0.9	7.7
02H	Handbrake operations	41	0.9	177	0.5	4.3
20E	Track–train interaction (hunting) (car)	40	0.9	419	1.1	10.5
05E	Other brake defect (car)	37	0.9	187	0.5	5.1
08E	Truck structure defects (car)	35	0.8	265	0.7	7.6
07H	Switching rules	30	0.7	198	0.5	6.6
02E	Brake rigging defect (car)	27	0.6	148	0.4	5.5
01E	Air hose defect (car)	19	0.4	148	0.4	7.8
01S	Signal failures	17	0.4	121	0.3	7.1
17E	All other locomotive defects	13	0.3	155	0.4	11.9
11T	Turnout defects: frogs	11	0.3	97	0.3	8.8
08H	Mainline rules	11	0.3	56	0.1	5.1
16E	Locomotive electrical and fires	10	0.2	28	0.1	2.8
04E	UDE (car or locomotive)	8	0.2	86	0.2	10.8
03H	Brake operations (other)	4	0.1	47	0.1	11.8
05H	Failure to obey or display signals	4	0.1	23	0.1	5.8
04H	Employee physical condition	3	0.1	41	0.1	13.7
06H	Radio communications error	3	0.1	13	0.0	4.3
14E	TOFC–COFC defects	2	0.0	2	0.0	1.0
03E	Handbrake defects (car)	1	0.0	2	0.0	2.0
	Total	4,352	100	37,456	100	8.6

NOTE: UDE = undesired emergency (brake application); TOFC = trailer on flat car; COFC = container on flat car.

(Spearman  $\rho = 0.95$ ,  $P < .01$ ). Certain derailment causes, notably broken rails or welds, are the most frequent when using either metric; consequently, efforts to prevent these high-frequency, high-severity accidents receive considerable attention.

Derailment frequency and severity (average number of cars derailed) were plotted against one another, with frequency on the abscissa and severity on the ordinate (Figure 1). The graph is divided into four quadrants on the basis of the average derailment frequency and severity along each axis. The graph enables easy comparison of the relative frequency and severity of different causes. Those causes in the upper right quadrant are most likely to pose the greatest risk because they are both more frequent and more severe than the average. The five cause groups are

- Broken rails or welds,
- Wide gauge,
- Buckled track,
- Obstructions, and
- Main-line brake operation.

Four other cause groups that are notable because of their high frequency of occurrence are

- Track geometry (excluding wide gauge),
- Bearing failure (car),

- Broken wheels (car), and
- Train handling (excluding brakes).

Three other cause groups are notable because of the high average severity of the resultant derailments and because they all have related causes:

- Rail defects at bolted joints,
- Other rail and joint defects, and
- Joint bar defects.

These three causes, along with the related cause group, broken rails or welds, are of particular interest, because when combined they accounted for almost 20% of all derailments and more than 30% of all derailed cars on Class I main lines (Table 2).

### Derailments on Siding and Yard Tracks

As discussed above, main track derailments are likely to be the most serious, but understanding the causes of derailments on siding and yard tracks is worthwhile because certain causes and solutions may apply to both. A chi-square test was conducted to compare the distributions of derailment frequency by the top 20 main-line derailment causes on main, yard, and siding tracks. The distribution of

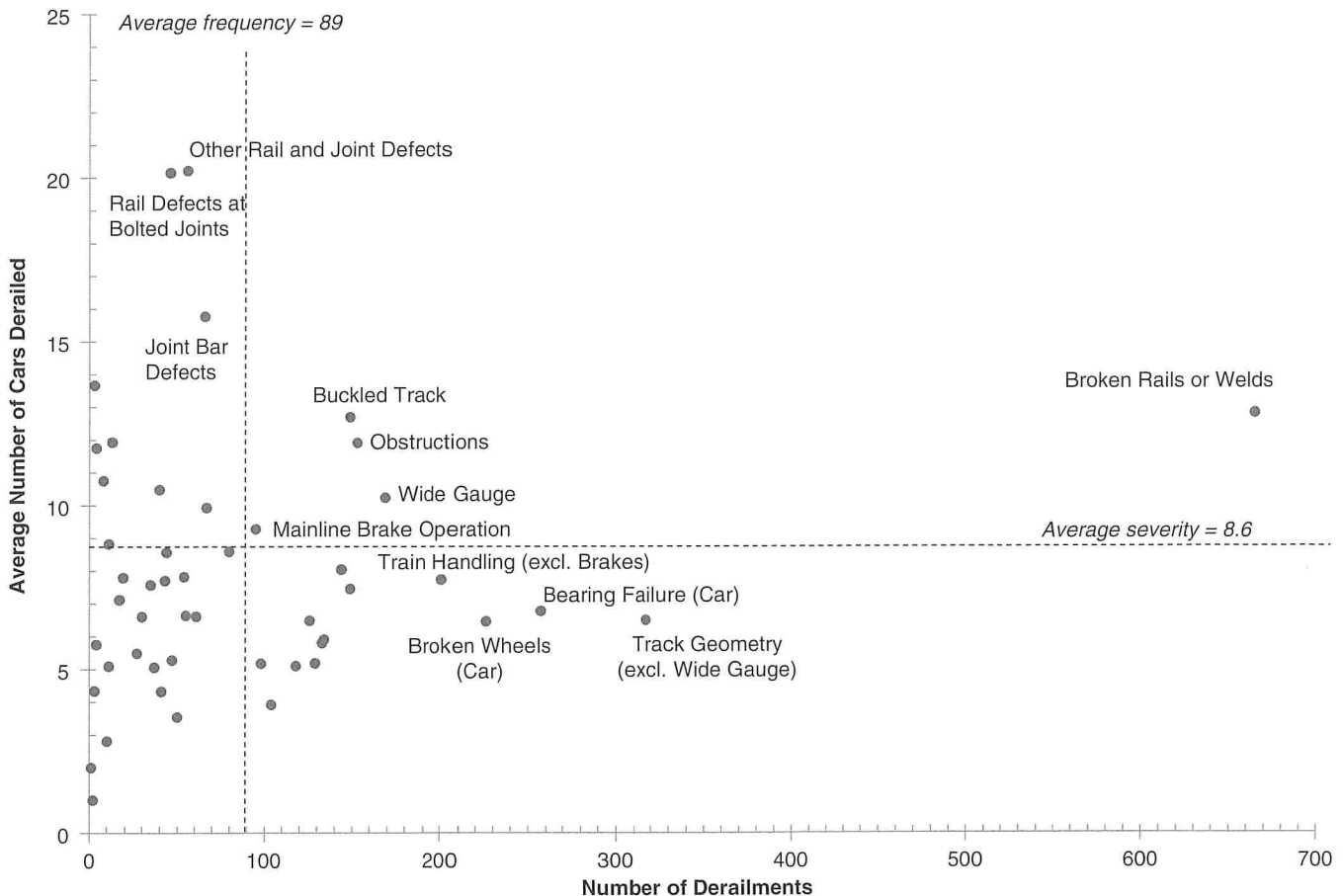


FIGURE 1 Frequency and severity graph of Class I main-line freight train derailments, 2001–2010.

TABLE 3 Top 10 Accident Causes of Freight Train Derailments by Track Type: Number of Derailments

Freight Train Derailments						
Rank	Main		Siding		Yard	
	Cause Group	Percentage	Cause Group	Percentage	Cause Group	Percentage
1	Broken rails or welds	15.3	Broken rails or welds	16.5	Broken rails or welds	16.4
2	Track geometry (excluding wide gauge)	7.3	Wide gauge	14.2	Use of switches	13.5
3	Bearing failure (car)	5.9	Turnout defects—switches	9.7	Wide gauge	13.5
4	Broken wheels (car)	5.2	Switching rules	7.7	Turnout defects—switches	11.1
5	Train handling (excluding brakes)	4.6	Track geometry (excluding wide gauge)	7.2	Train handling (excluding brakes)	6.7
6	Wide gauge	3.9	Use of switches	5.8	Switching rules	6.2
7	Obstructions	3.5	Train handling (excluding brakes)	3.5	Track geometry (excluding wide gauge)	3.6
8	Buckled track	3.4	Lading problems	2.3	Miscellaneous track and structure defects	3.4
9	Track–train interaction	3.4	Roadbed defects	2.1	Track–train interaction	3.1
10	Other axle or journal defects (car)	3.3	Miscellaneous track and structure defects	2.1	Other miscellaneous	3.0

derailment causes varied significantly by track type ( $\chi^2 = 1,780$ ,  $df = 38$ ,  $P < .01$ ). Tables 3 and 4 show the top 10 accident causes by derailment frequency and total number of cars derailed, respectively, on different track types.

Comparison of main tracks with yard and siding tracks using either metric, derailment frequency or number of cars derailed, reveals that broken rails or welds were the leading derailment cause on all

three track types. However, by contrast to main tracks, bearing failures and broken wheels were not among the top accident causes on yard and siding tracks, probably because of lower operating speeds. Instead, human factor–related causes such as improper use of switches and violation of switching rules were more prevalent. Misaligned switches caused 14% of yard derailments, and this cause has received particular attention in recent years. The higher incidence

TABLE 4 Top 10 Accident Causes of Freight Train Derailments by Track Type: Number of Cars Derailed

Freight Cars Derailed Because of Train Derailments						
Rank	Main		Siding		Yard	
	Cause Group	Percentage	Cause Group	Percentage	Cause Group	Percentage
1	Broken rails or welds	22.7	Broken rails or welds	23.2	Broken rails or welds	19.3
2	Track geometry (excluding wide gauge)	5.5	Wide gauge	13.8	Wide gauge	18.2
3	Buckled track	5.0	Turnout defects—switches	10.4	Use of switches	10.0
4	Obstructions	4.9	Track geometry (excluding wide gauge)	6.2	Turnout defects—switches	9.8
5	Bearing failure (car)	4.6	Use of switches	4.8	Train handling (excluding brakes)	7.7
6	Wide gauge	4.6	Switching rules	4.0	Miscellaneous track and structure defects	4.2
7	Train handling (excluding brakes)	4.1	Train handling (excluding brakes)	3.5	Switching rules	3.9
8	Broken wheels (car)	3.9	Obstructions	3.0	Track geometry (excluding wide gauge)	3.3
9	Other axle or journal defects (car)	3.1	Buckled track	2.8	Track–train interaction	3.2
10	Other rail and joint defects	3.0	Brake operation (main line)	2.7	Brake operation (main line)	2.7



of switch-related derailments in yards and sidings compared with main lines is probably due to the greater number and more frequent use of turnouts on these tracks and thus the greater likelihood of error. Another consequence of the more frequent use of switches is the greater prevalence of switch defects. Switch defects caused approximately 10% of derailments on yard and siding tracks, but only 3% on main lines. The reason for this is probably twofold: the more frequent use of switches on these types of tracks subjects them to greater exposure and thus more opportunity to cause a derailment, and because of their heavy use, the switches are subject to more wear and tear and consequently faster deterioration. The switch points are typically the most vulnerable parts of switches, so their protection and lubrication, along with improved wheel profile and truck steering performance, may offer means to prevent switch-defect derailments (26–28). Another difference between main tracks compared with yard and siding tracks is that wide gauge accounted for 14% of derailments on siding and yard tracks but for only 4% on main lines. Again, this difference is probably due to the lower speed characteristic of yard and siding tracks but with a different explanation; the lower operating speed permits greater tolerances in the

track gauge standards, and therefore these tracks may be more prone to this type of derailment (29).

### EFFECT OF DERAILMENT SPEED

So far this paper has considered accident and track type as factors affecting the likelihood that a train or rail car will derail, but another important parameter affecting derailments is train speed at the time of derailment. Indeed, speed may be a contributing factor to some of the differences cited above. Common sense demands that speed is a factor affecting derailment severity, and previous research has established several qualitative and quantitative relationships between derailments and speed (1). The top 10 accident causes of main-line train derailments were sorted into different groups, corresponding to the FRA track class speed ranges, and compared, again by derailment frequency and number of cars derailed (Figure 2) (29).

In all speed ranges, broken rails or welds were the leading cause of derailments; however, the relative frequency of the next most common accident types differed substantially for lower versus

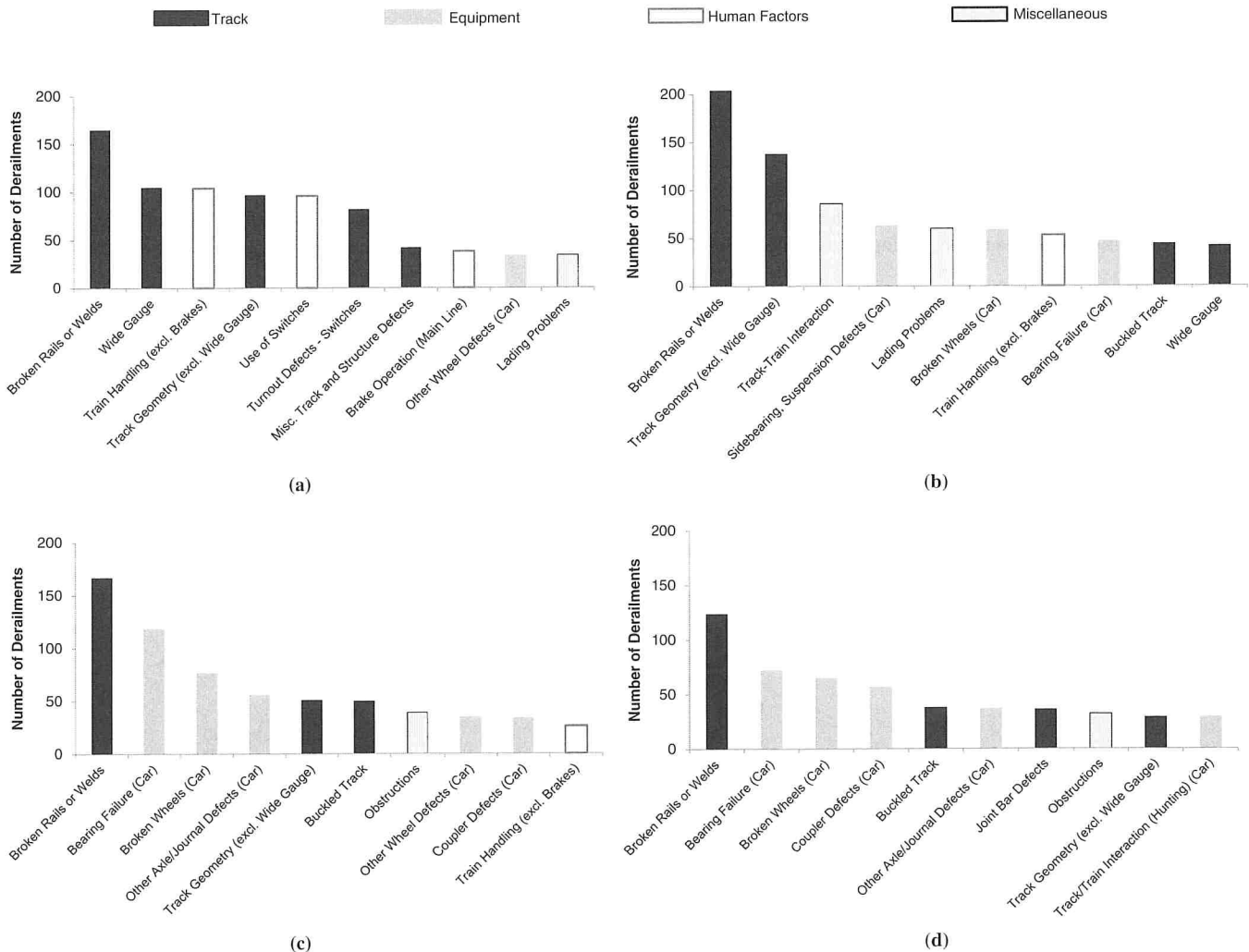


FIGURE 2 Number of freight train derailments by speed and accident cause on Class I main lines: (a) derailment speed 0–10 mph, (b) derailment speed 10–25 mph, (c) derailment speed 25–40 mph, and (d) derailment speed 40–80 mph.

higher speed derailments. At speeds below 10 mph, certain track-related and human factor-related causes occurred more frequently than equipment-related causes. But at derailment speeds greater than 25 mph, human factors accidents such as improper train handling, braking operations, and improper use of switches were almost completely absent, replaced by equipment causes, such as bearing failure, broken wheel, and axle and journal defects. The derailment frequency distribution for 49 main-line accident causes and three derailment-speed groups (<10 mph, 11–25 mph, and >25 mph) were compared in a chi-square analysis, and the results were significant ( $\chi^2 = 1,192$ ,  $df = 96$ ,  $P < .01$ ), indicating an association between accident cause and derailment speed.

## ACCIDENT PREVENTION STRATEGY

To gain insights into the potential safety benefits of strategies to reduce various types of derailments, a sensitivity analysis was conducted (Figure 3, *a* and *b*). An estimation was done to determine by what percentage main track train and car derailment rates would be reduced in the event that certain accident causes were reduced or eliminated. Four of the leading main-line accident causes were considered: broken rails or welds (08T), track geometry defects (04T), bearing failure (10E), and broken wheels (12E). There are a number of approaches in practice or being developed that may address these. Broken rail preventive measures include rail inspection, rail grinding, rail repair,

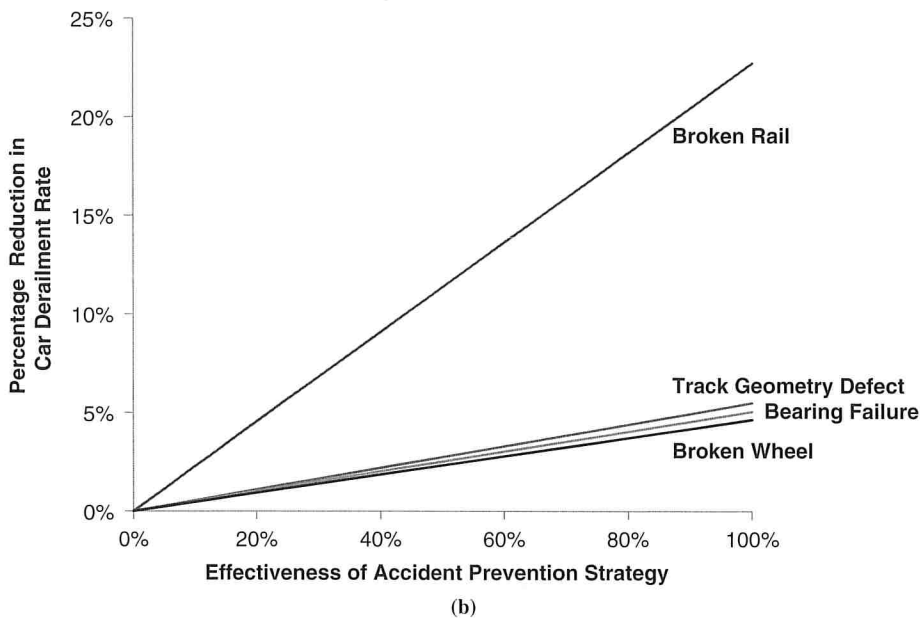
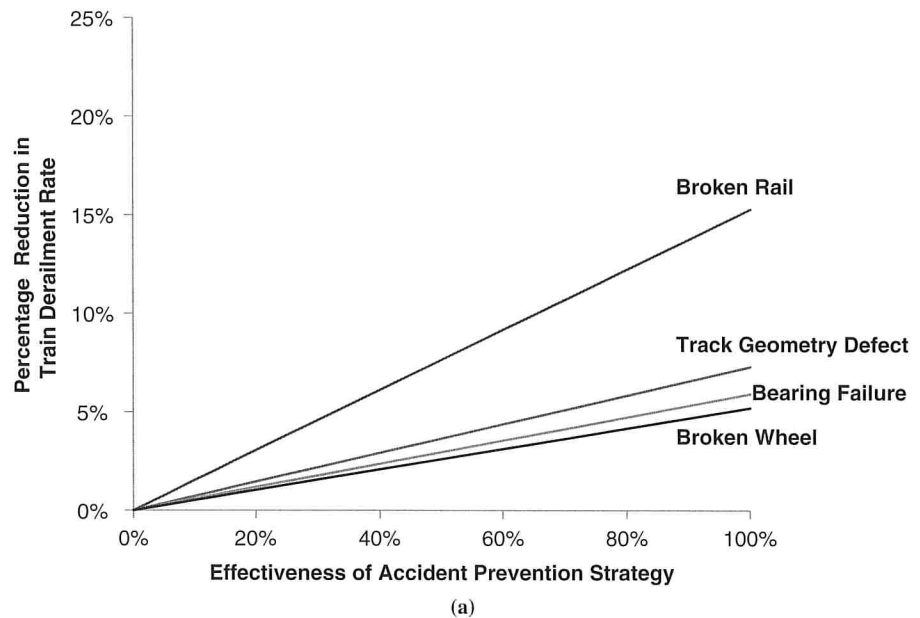


FIGURE 3 Percentage reduction in derailment rates by derailment prevention strategy: (a) train derailment rate reduction and (b) car derailment rate reduction.

and renewal (30). Track geometry inspection and maintenance are often based on some measurable indices, such as the track quality index (31, 32). Wayside and onboard detection systems aim to identify and inform railroads and car owners about the need to remove or repair rolling stock defects before they cause an accident. Hot bearing detectors and wheel impact load detectors are commonly used to detect problems with those components (33, 34). Nevertheless, these techniques and technologies are not 100% effective in eliminating all the accident causes they are intended to prevent, and research and development are ongoing to further develop their capability. The effectiveness of an accident prevention strategy is defined here as the percentage of the maximum safety benefit it might potentially realize. The sensitivity analysis helps illustrate the relevant potential safety benefit that might be realized if technologies or techniques were implemented with varying degrees of effectiveness. For example, broken rails or welds caused 15.3% of train derailments and 22.7% of cars derailed (Table 2); thus, if all broken-rail-related causes were eliminated, train and car derailment rates would decline by a corresponding amount. Even if only 50% of broken rails or welds could be prevented, the prevention would result in a larger percentage reduction in train and car derailment rates than would any of the other accident prevention strategies at 100% effectiveness.

The effects of different accident prevention strategies may not necessarily be independent of one another. For example, improved wheel condition can reduce dynamic loading of track, thereby reducing track defect rates, and vice versa. The interactive damage forces between track and equipment have been discussed in previous studies (35, 36). Resor and Zarembski proposed an engineering model to estimate the change in relative damage to track and equipment given the change in impact load (35). With the use of their model, it was estimated that a 1% reduction in impact load would result in a 1.3% reduction in damage to track and a 0.6% reduction in damage to wheels and axles. Nevertheless, the authors are unaware of any research that quantifies the reduction in track and equipment damages, with the corresponding reduction in accident probability. Further research is needed to understand better what the possible interactive effects are, how to quantify them where they exist, and what their effects are on accident rate estimation and safety policy evaluation. For purposes of illustration, independence of individual derailment prevention strategies was assumed in the sensitivity analyses presented here. To the extent that interactive effects among different accident prevention strategies reduce the safety benefits due to another, the analyses here may slightly overstate the benefit of a particular derailment cause prevention measure if other related measures were implemented.

This paper focuses on developing an analytical framework to understand the relative importance of different accident causes under various operating conditions. The analyses presented here are just the first step in a risk-based approach to derailment prevention. The implementation costs of different risk reduction measures may be affected by the effectiveness of technology, extent of implementation, installation and maintenance practices, and many other factors. Schafer and Barkan estimated \$900 per track mile as the annual cost for ultrasonic and geometric track inspection and \$1,900 per track mile for rail grinding on one Class I railroad (30). Robert et al. reported a total cost of \$86 million for implementing wayside detectors in the United States from 1994 to 2008 (37). However, the proportion of the costs directly related to safety improvement is not well understood and further study is required. An additional complexity is that safety improvement activities may affect operational efficiency differently in different time periods. For example, track maintenance may cause

train delay in the short term but improve efficiency in the long term by reducing the potential service disruptions due to accidents.

Further research is needed to understand the relationship between accident rates and occurrence in regard to accident frequency and corresponding traffic exposure. This research will enable a better comparison of the accident risk under different operating conditions, such as main lines versus yard tracks. The next step is to quantify the benefits and costs of specific risk reduction measures, thereby allowing integration of the multiple trade-offs involving safety, efficiency, and cost. In that way, interactive effects between strategies can be accounted for, and the optimal combination of investment strategies selected for any given level of financial resources.

## CONCLUSIONS

Accident cause distribution varies by accident type, track type, and speed. Derailments are the most common type of train accident on each track type, and the majority of cars derailed are due to train derailments. Track and equipment failures are the primary causes of train derailments on main tracks, whereas the use of switches and switching rules has a substantial effect on derailment frequency on siding and yard tracks. Some accident causes tend to occur more frequently at higher speeds, whereas others are more likely at lower speeds. The interactive effects of derailment speed and accident cause affect train accident frequency and severity.

The safety benefits of accident prevention strategies were evaluated according to the percentage reduction in train and car derailment rates. Prevention of broken rails or welds is expected to yield a larger percentage reduction in train and car derailment rates than other accident prevention strategies. However, the cost-effectiveness of this and other accident prevention strategies must be properly compared to select the most efficient means of improving railroad train operating safety. Ultimately these strategies should be considered as part of an integrated framework to optimize investment that maximizes safety benefits and minimizes risk.

## ACKNOWLEDGMENTS

The first author was supported in part by grants from BNSF Railway and ABSG Consulting. The authors are grateful to Donald Bullock and Laura Ghosh of the University of Illinois for their comments on the revised manuscript. The authors greatly appreciate the useful comments from four anonymous reviewers. This research was partially supported by a grant from the NEXTRANS University Transportation Center.

## REFERENCES

1. Barkan, C. P. L., C. T. Dick, and R. Anderson. Railroad Derailment Factors Affecting Hazardous Materials Transportation Risk. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 1825, Transportation Research Board of the National Academies, Washington, D.C., 2003, pp. 64–74.
2. Anderson, R. T. *Quantitative Analysis of Factors Affecting Railroad Accident Probability and Severity*. MS thesis. University of Illinois at Urbana–Champaign, 2005.
3. Liu, X., C. P. L. Barkan, and M. R. Saat. Analysis of Derailments by Accident Cause: Evaluating Railroad Track Upgrades to Reduce Transportation Risk. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 2261, Transportation Research Board of the National Academies, Washington, D.C., 2011, pp. 178–185.

4. *FRA Guide for Preparing Accident/Incident Reports*. FRA, U.S. Department of Transportation, 2011.
5. Chadwick, S., M. R. Saat, and C. P. L. Barkan. Analysis of Factors Affecting Train Derailment at Highway–Rail Grade Crossings. Presented at 91st Annual Meeting of the Transportation Research Board, Washington, D.C., 2012.
6. Association of American Railroads. *Class I Railroad Statistics*. <http://www.aar.org/-/media/aar/Industry%20Info/AAR%20Stats%202010%201123.ash>.
7. Saccomanno, F. F., J. H. Shortreed, and M. Van Aerde. *Assessing the Risks of Transporting Dangerous Goods by Truck and Rail*. Institute for Risk Research, University of Waterloo, Waterloo, Ontario, Canada, 1988.
8. Austin, R. D., and J. L. Carson. An Alternative Accident Prediction Model for Highway–Rail Interfaces. *Accident Analysis and Prevention*, Vol. 34, 2002, pp. 31–42.
9. Arthur D. Little, Inc. (ADL). *Risk Assessment for the Transportation of Hazardous Materials by Rail, Supplementary Report: Railroad Accident Rate and Risk Reduction Option Effectiveness Analysis and Data*, 2nd rev. ADL, Cambridge, Mass., 1996.
10. Schafer, D. H., II, and C. P. L. Barkan. Relationship Between Train Length and Accident Causes and Rates. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 2043, Transportation Research Board of the National Academies, Washington, D.C., 2008, pp. 73–82.
11. Yang, T. H., W. P. Manos, and B. Johnstone. *A Study Continuation of Derailment Behavior Final Report (Phase 08 Report on Computer Derailment Study)*. RPI/AAR Report RA-08-1-12 (R-135). Railroad Tank Car Safety Research and Test Project. Association of American Railroads, Washington, D.C., 1972.
12. Yang, T. H., W. P. Manos, and B. Johnstone. Dynamic Analysis of Train Derailments. 72-WA/RT-6. *Rail Transportation Proceedings*. The American Society of Mechanical Engineers, New York, 1973, p. 8.
13. Nayak, P. R., D. B. Rosenfield, and J. H. Hagopian. *Event Probabilities and Impact Zones for Hazardous Materials Accidents on Railroads*. Report DOT/FRA/ORD-83/20. FRA, U.S. Department of Transportation, 1983.
14. Glickman, T. S., and D. B. Rosenfield. Risks of Catastrophic Derailments Involving the Release of Hazardous Materials. *Management Science*, Vol. 30, No. 4, 1984, pp. 503–511.
15. Coppens, A. J., J. D. E. Wong, A. Bibby, A. M. Birk, and R. J. Anderson. *Development of a Derailment Accident Computer Simulation Model*. Transport Canada Report No. TP 9254E. Prepared for the Transportation Development Centre and Transport of Dangerous Goods, Ottawa, Ontario, Canada, 1988.
16. Saccomanno, F. F., and S. El-Hage. Minimizing Derailments of Railcars Carrying Dangerous Commodities Through Effective Marshaling Strategies. In *Transportation Research Record 1245*, TRB, National Research Council, Washington, D.C., 1989, pp. 34–51.
17. Treichel, T. T., and C. P. L. Barkan. *Working Paper on Mainline Freight Train Accident Rates*. Research and Test Department, Association of American Railroads, Washington, D.C., 1993.
18. Dick, C. T., C. P. L. Barkan, E. R. Chapman, and M. P. Stehly. Multivariate Statistical Model for Predicting Occurrence and Location of Broken Rails. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 1825, Transportation Research Board of the National Academies, Washington, D.C., 2003, pp. 48–55.
19. Anderson, R. T., and C. P. L. Barkan. Railroad Accident Rates for Use in Transportation Risk Analysis. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 1863, Transportation Research Board of the National Academies, Washington, D.C., 2004, pp. 88–98.
20. Anderson, R. T., and C. P. L. Barkan. Derailment Probability Analyses and Modeling of Mainline Freight Trains. *Proc., 8th International Heavy Haul Railway Conference*, International Heavy Haul Association, Rio de Janeiro, Brazil, 2005.
21. Saat, M. R., and C. P. L. Barkan. *Tank Car Safety Design vs. Infrastructure Improvements in Reducing Hazardous Materials Transportation Risks*. Presented at INFORMS Annual Meeting, Pittsburgh, Pa., 2006.
22. Liu, X., M. R. Saat, and C. P. L. Barkan. Benefit–Cost Analysis of Infrastructure Improvement for Derailment Prevention. *Proc., ASME-IEEE-ASCE-AREMA-TRB Joint Rail Conference* (CD-ROM), University of Illinois at Urbana-Champaign, 2010.
23. Kawprasert, A. *Quantitative Analysis of Options to Reduce Risk of Hazardous Materials Transportation by Railroad*. PhD dissertation. University of Illinois at Urbana-Champaign, Urbana, 2010.
24. English, G. W., G. Higham, M. Bagheri, T. W. Moynihan, and F. F. Saccomanno. *Evaluation of Risk Associated with Stationary Dangerous Goods Railway Cars*. Transport Canada Report No. TP 14690E. Prepared for the Transportation Development Centre (TDC), Montreal, Quebec, Canada, 2007.
25. Liu, X., C. P. L. Barkan, and M. R. Saat. *Probability Analysis of Hazardous Materials Releases in Railroad Transportation*. Presented at INFORMS Annual Meeting, Austin, Tex., 2010.
26. Zaremski, A. M. *Derailment of Transit Vehicles in Special Trackwork*. Transit Cooperative Research Program, 1997.
27. Wolf, G. Switch Point Derailments: Is It the Point or the Wheel? *Interface, Journal of Wheel–Rail Interaction*, July 2006.
28. Wu, H., and N. Wilson. Railway Vehicle Derailment and Prevention. In *Handbook of Railway Vehicle Dynamics*, Taylor and Francis Group, Boca Raton, Fla., 2006.
29. *Track Safety Standards*. FRA. 49 CFR 213, 2003.
30. Schafer, D. H., and C. P. L. Barkan. A Prediction Model for Broken Rails and an Analysis of Their Economic Impact. *Proc., American Railway Engineering and Maintenance of Way Association (AREMA) Annual Conference*, Salt Lake City, Utah, Sept. 2008.
31. Uzarski, D. R. *Development of a Track Structure Condition Index*. PhD dissertation. University of Illinois at Urbana-Champaign, 1991.
32. El-Sibaie, M., and Y.-J. Zhang. Objective Track Quality Indices. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 1863, Transportation Research Board of the National Academies, Washington, D.C., 2004, pp. 81–87.
33. Schlake, B. *Impact of Automated Condition Monitoring Technologies on Railroad Safety and Efficiency*. MS thesis. University of Illinois at Urbana-Champaign, 2010.
34. Kalay, S., P. French, and H. Tournay. The Safety Impact of Wagon Health Monitoring in North America. *Proc., 9th World Congress on Railway Research Conference*, World Congress on Railway Research, Lille, France, 2011.
35. Resor, R. R., and A. M. Zaremski. Factors Determining the Economics of Wayside Defect Detectors. Presented at 83rd Annual Meeting of the Transportation Research Board, Washington, D.C., 2004.
36. Bladon, K., D. Rennison, G. Lzbinsky, R. Tracy, and T. Bladon. Predictive Condition Monitoring of Railway Rolling Stock. *Proc., Conference on Railway Engineering*, Darwin, Australia, June 2004.
37. Robert, W., A. Aeppli, and P. Little. *Post-Audit of Wayside Detector Costs and Benefits*. Cambridge Systematics Inc., Cambridge, Mass., Sept. 2009.

*The Railroad Operational Safety Committee peer-reviewed this paper.*





**TRANSPORTATION RESEARCH RECORD:  
JOURNAL OF THE TRANSPORTATION RESEARCH BOARD**

## **Peer Review Process**

The *Transportation Research Record: Journal of the Transportation Research Board* publishes approximately 25% of the more than 4,000 papers that are peer reviewed each year. The mission of the Transportation Research Board (TRB) is to disseminate research results to the transportation community. The Record series contains applied and theoretical research results as well as papers on research implementation.

The TRB peer review process for the publication of papers allows a minimum of 30 days for initial review and 60 days for rereview, if needed, to ensure that only the highest-quality papers are published. A minimum of three reviews are required for a publication recommendation. The process also allows for scholarly discussion of any paper scheduled for publication, along with an author-prepared closure.

The basic elements of the rigorous peer review of papers submitted to TRB for publication are described below.

### **Paper Submittal: June 1–August 1**

Papers may be submitted to TRB at any time. However, most authors use the TRB web-based electronic submission process available between June 1 and August 1, for publication in the following year's Record series.

### **Initial Review: August 15–November 15**

TRB staff assigns each paper by technical content to a committee that administers the peer review. The committee chair assigns at least three knowledgeable reviewers to each paper. The initial review is completed by mid-September.

By October 1, committee chairs make a preliminary recommendation, placing each paper in one of the following categories:

1. Publish as submitted or with minor revisions;
2. Reconsider for publication, pending author changes and re-review; or
3. Reject for publication.

By late October, TRB communicates the results of the initial review to the corresponding author. Corresponding authors communicate the information to coauthors. Authors of papers in Category 2 (above) must submit a revised version addressing all reviewer comments, along with an explanation of how the comments have been addressed.

### **Rereview: November 20–January 25**

The committee chair reviews revised papers in Category 1 (above) to ensure that the changes are made and sends the Category 2 revised papers to the initial reviewers for rereview. After rereview, the chair makes the final recommendation on papers in Categories 1 and 2. If the paper has been revised to the committee's satisfaction and ranks among the best papers, the chair may recommend publication. The chair communicates the results of the rereview to the authors.

### **Discussions and Closures: February 1–May 15**

Discussions may be submitted for papers that will be published. TRB policy is to publish the paper, the discussion, and the author's closure in the same Record.

Many papers considered for publication in the *Transportation Research Record* are also considered for presentation at TRB meetings. Individuals interested in submitting a discussion of any paper presented at a TRB meeting must notify TRB no later than February 1. If the paper has been recommended for publication in the *Transportation Research Record*, the discussion must be submitted to TRB no later than April 15. A copy of this communication is sent to the author and the committee chair.

The committee chair reviews the discussion for appropriateness and asks the author to prepare a closure to be submitted to TRB by May 15. The committee chair reviews the closure for appropriateness. After the committee chair approves both discussion and closure, the paper, the discussion, and the closure are included for publication together in the same Record.

### **Final Manuscript Submittal: March 15**

In early February, TRB requests a final manuscript for publication—to be submitted by March 15—or informs the author that the paper has not been accepted for publication. All accepted papers are published by December 31.

### **Paper Awards: April to January**

The TRB Executive Committee has authorized annual awards sponsored by Groups in the Technical Activities Division for outstanding published papers:

- Charley V. Wootan Award (Policy and Organization Group);
- Pyke Johnson Award (Planning and Environment Group);
- K. B. Woods Award (Design and Construction Group);
- D. Grant Mickle Award (Operations and Preservation Group);
- John C. Vance Award (Legal Resources Group);
- Patricia F. Waller Award (Safety and System Users Group); and
- William W. Millar Award (Public Transportation Group).

Other Groups also may nominate published papers for any of the awards above. In addition, each Group may present a Fred Burggraf Award to authors 35 years of age or younger.

Peer reviewers are asked to identify papers worthy of award consideration. Each Group reviews all papers nominated for awards and makes a recommendation to TRB by September 1. TRB notifies winners of the awards, which are presented at the following TRB Annual Meeting.

**Transportation Research Board**  
**www.TRB.org**

**TRANSPORTATION RESEARCH BOARD**

500 Fifth Street, NW  
Washington, DC 20001

ADDRESS SERVICE REQUESTED

NON-PROFIT ORG.  
U.S. POSTAGE  
PAID  
WASHINGTON, D.C.  
PERMIT NO. 8970



**THE NATIONAL ACADEMIES™**

*Advisers to the Nation on Science, Engineering, and Medicine*

The nation turns to the National Academies—National Academy of Sciences, National Academy of Engineering, Institute of Medicine, and National Research Council—for independent, objective advice on issues that affect people's lives worldwide.

[www.national-academies.org](http://www.national-academies.org)

ISBN 978-0-309-22329-4

90000



9 780309 223294